

# nature

## Nature at 150: evidence in pursuit of truth

A century and a half has seen momentous changes in science. But evidence and transparency are more important than ever before.

made its way into the world. Its ambition was intellectually bold and commercially risky: to bring news of the latest discoveries and inventions to scientists and the public alike. Although the journal was aimed at a broad audience, scientists took a particular liking to it – because it allowed them to communicate their findings quickly. Nature's weekly schedule offered a refreshing contrast to the leisurely timescales of learned-society journal publishing and conference proceedings. And, as universities grew, more 'letters to the editor' from scientists started arriving at the Nature offices in London. The journal became a venue for publishing discoveries because its writers also became its readers – and we have been trying to serve scientists and society ever since.

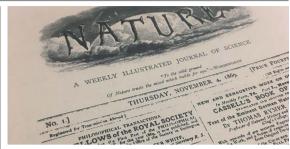
n 4 November 1869, the first issue of Nature

In this, Nature's 150th anniversary issue, we're celebrating and remembering many of the notable discoveries that authors have communicated in the journal's pages, along with the agenda-setting journalism and commentary that has always been an essential part of our voice.

A century and a half is long enough to see how our understanding of the natural world changes with each instalment of new evidence. Take human origins. In February 1925, Nature published the discovery by Raymond Dart of Australopithecus africanus in South Africa<sup>1</sup>. It was the first fossil link between humans and apes, and it caused a sensation, providing evidence that humans evolved from a common ancestor in Africa as Charles Darwin had proposed – and not in Britain or Indonesia as had previously

Nearly 80 years later, the discovery of the remains of Homo floresiensis in 2004, which came to be known as the hobbit, demonstrated that our genus was remarkably diverse<sup>2</sup>. Further revelations about human prehistory and evolution quickly followed, culminating in advances in ancient genomics. These have revealed that, as recently as 30,000 to 60,000 years ago, humans coexisted and had offspring with other hominins - Neanderthals and Denisovans3.

Nature also published some of the remarkable developments that took place in physics in the early part of the twentieth century. These include James Chadwick's proposal in 1932 of the existence of a new particle, the Researchers and their remarkable discoveries have made us what we are."



Nature made its debut on 4 November 1869.

neutron, to add to the electron and the proton<sup>4</sup>. Today, many more fundamental particles have been discovered because of the predictions of the standard model of particle physics. Some of the earliest findings of exoplanets appeared in our pages, including, in 1995, the first report of an exoplanet orbiting around a Sun-like star in another galaxy<sup>5</sup> – for which Michel Mayor and Didier Queloz won a share of the 2019 Nobel Prize in Physics.

Arguably, Nature's most memorable publications were the reports in April 1953 on the structure of DNA – including papers from Maurice Wilkins<sup>6</sup> and Rosalind Franklin<sup>7</sup>, in addition to the paper by Francis Crick and James Watson<sup>8</sup>. The discovery that DNA was a double helix changed biology forever. Forty years later, we proudly published the first draft sequence of a human genome carried out by a publicly funded group, the International Human Genome Sequencing Consortium<sup>9</sup>. Without the researchers' collective achievement, medicine, agriculture, conservation and criminal justice would look very different today.

There is, of course, no definitive list of the most influential or important pieces of research that Nature has published. A series of News & Views articles on page 35 explains the importance and lasting impact of ten key papers from our archive. We also chose a long list of 150 interesting, illuminating, entertaining and sometimes controversial articles - one for every year of our life – and have been posting one per day on social media for the past few months. But even compiling this longer list involved vigorous and sometimes tense discussion among the editors.

At the start of the year, we also began discussing what to feature on our anniversary issue cover. The result – a data analysis of Nature's archive which highlights the multidisciplinary scope of the journal - is rendered as the extraordinary fireworks you can see on the cover and in a video and interactive visualization online. Our anniversary issue includes a rich variety of written and multimedia content on the past, present and future of Nature and of research itself.

#### Responsible science

As science has advanced during the past century and a half, discovery has gone hand in hand with world-changing inventions – particularly in industrial-scale technologies. Many of these technologies, from the internal combustion engine to synthetic agrochemicals, have improved the quality of life for hundreds of millions of people; but at

## nature

the same time they have also damaged the environment or raised serious ethical and safety concerns.

In some cases, researchers have been able to sound the alarm in time for remedial action, as chemists Mario Molina and Sherwood Rowland did in June 1974 when they worked out that chlorine originating from chlorofluorocarbons (CFCs) was destroying atmospheric ozone 10. A decade later, physicist Joe Farman and colleagues showed that ozone levels over Antarctica were lower than expected — the first detection of the ozone hole 11.

These findings led to the 1989 Montreal Protocol, an international agreement to cut ozone-depleting substances, and a shining example of how people can unite to take action when scientific evidence points to an impending environmental disaster. Sadly, the same cannot yet be said for climate change, even though researchers have been sounding ever-louder warnings since the 1970s that greenhouse-gas emissions are warming the planet.

As the pace of discovery and invention accelerates — from isolating stem cells  $^{12}$  to the development of cloning  $^{13}$  and gene-editing technologies, to last month's description of quantum supremacy  $^{14}$  — there is a clear need, perhaps now more than ever, for researchers and research publishers to acknowledge, and implement, our responsibility to society. We must commit to greater openness and ensure that findings are reproducible, and we must act with integrity at all times. *Nature* and the researchers it serves have a duty to work side by side with those in our broader society who will be affected by the products of research, and to consider generations to come.

#### **Room to improve**

Looking back, there have been times when *Nature* did not adhere to standards that we hold ourselves to today. We should have called out when Jocelyn Bell Burnell was overlooked for the Nobel physics prize for her work in the discovery of pulsars<sup>15</sup>. And it shouldn't have taken until 2007 for us to replace the phrase "scientific men" with "scientists" in our mission statement.

Organized peer review – the cornerstone of scientific publishing – was introduced in *Nature* only after 1966, although we have tried to make up for lost time since. In 2006, *Nature* conducted trials on open peer review; we now offer authors double-blind peer review, and are one of several journals in the Nature family to offer reviewers the opportunity to be named.

Another area where overdue change is under way is in the people represented in the journal. In the early years, *Nature* was dominated by papers with one or two authors, mostly male, and mostly from the Northern Hemisphere. Today, papers with a single author are almost unheard of and author lists can run to the thousands, reflecting the increasingly team-based nature of current research. Although most of our authors still come from institutions in Europe and North America — where most research funding is concentrated — our author community is becoming more geographically diverse.

But researchers from large parts of the world, notably Africa, remain under-represented. This reflects broader

The most exciting and dramatic changes will be the ones we cannot imagine today."

inequalities that stem from the uncomfortable historical reality that science and empire often worked in a symbiotic relationship. We recognize that *Nature* was founded at the height of such an age. Change will need time, but we are committed to doing more to make a difference.

#### View to the future

As the boundaries between disciplines blur and research becomes increasingly multi- and transdisciplinary, *Nature* is moving beyond a traditional focus on the natural sciences to embrace social sciences, translational and clinical research and applied science and engineering. Looking to the future, we hope to contribute to greater transparency and openness in academia. We will probably see even more collaborative ways of doing research and more changes in the way it is published.

Predicting the future is notoriously difficult. Writer William Gibson, in his 1984 cyberpunk novel *Neuromancer*, foresaw a form of today's stem-cell therapy and sophisticated artificial intelligence, but failed to anticipate mobile phones. Even in the early 1990s, relatively few people anticipated that 'electronic publishing', as it was starting to be called, would jeopardize the future of mass-produced printed journals. The most exciting and dramatic changes will be the ones we cannot imagine today.

It's unlikely that our founders imagined that, 150 years on, *Nature* would be publishing more than 850 research papers and 3,000 articles of news, opinion and analysis each year, and reaching around 4 million readers online each month. That's your doing: researchers and their remarkable discoveries have made us what we are. We have reached this important milestone only through listening, responding and adapting to the community we serve.

In other respects, *Nature* now is just the same as it was at the start. We will continue in our mission to stand up for research, serve the global research community and communicate the results of science around the world. We will strive to hold to account those in positions of responsibility in research, policy and industry, and to continue to advocate for fewer unintended harmful consequences of research for people and the planet.

Research, science, knowledge, scholarship – however we might choose to characterize the marshalling of evidence in the pursuit of truth – the values we hold are more important than ever before.

- 1. Dart, R. A. Nature 115, 195-199 (1925).
- 2. Brown P. et al. Nature 431, 1055-1061 (2004).
- 3. Reich, D. et al. Nature 468, 1053-1060 (2010).
- 4. Chadwick, D. Nature 129, 312 (1932).
- 5. Mayor, M. & Queloz, D. Nature 378, 355-359 (1995).
- 6. Wilkins, M. H. F., Stokes, A. R. & Wilson, H. R. Nature 171, 738–740 (1953).
- 7. Franklin, R. E. & Gosling, R. G. Nature 171, 740-741 (1953).
- 8. Watson, J. D. & Crick, F. H. C. *Nature* **171**, 737–738 (1953).
- International Human Genome Sequencing Consortium Nature 409, 860–921 (2001).
- 10. Molina, M. J. & Rowland, F. S. *Nature* **249**, 810–812 (1974)
- 11. Farman, J. C., Gardner, B. G. & Shanklin, J. D. Nature 315, 207-210 (1985).
- 12. Evans, M. J. & Kaufman, M. H. Nature 292, 154-156 (1981).
- Campbell, K. H. S., McWhir, J., Ritchie, W. A. & Wilmut, I. Nature 380, 64–66 (1996).
- 14. Arute, F. et al. Nature **574**, 505–510 (2019).
- Hewish, A., Bell, S. J., Pilkington, J. D. H., Scott, P. F. & Collins, R. A. Nature 217, 709–713 (1968).

# **World view**

## Better methods can't make up for mediocre theory



**Bv Paul Smaldino** 

With better questions, many reproducibility problems will fall away, says Paul Smaldino.

uch digital ink has been spilt describing ways to improve replicability in science. Preregistration. Open data. Open code. These are all necessary, but insufficient. The thing is, we don't just want science to be reproducible. We want it to help us to make better sense of the world. For that, we must create better hypotheses – and those require better models and better measurements.

A theoretical model of mine (P. E. Smaldino and R. McElreath Soc. Open Sci. 3, 160384; 2016) made headlines when it showed that bad science – or rather, less rigorous science that could produce more papers in less time - could crowd out the more robust sort. This suggested that generating better hypotheses is at least as important as reducing methodological errors for minimizing false discoveries.

Who cares if you can replicate an experiment that found that people think the room is hotter after reading a story about nice people? Will this help us to develop better theories? You can craft a fun story about that result, but can you devise the next great scientific question?

To generate good hypotheses, we need good theory. In a landmark study attempting to replicate 100 psychology papers, cognitive-psychology studies were replicated about twice as often as those from social psychology (Open Science Collaboration, Science 349, aac4716; 2015). I think that's because cognitive psychology has better theories.

Good theory has at least two requirements. First, it can be used to build mathematical or computational models that derive clear, testable consequences from our assumptions. Every mature scientific discipline has these. Physicists use models of force and momentum to predict the motion of materials. Epidemiologists use models of contagions to understand the spread of disease. Neuroscientists use models of neural-spike trains to understand information flow in the brain. Social scientists use game models to understand the emergence of social norms.

Second, good theory must make sense, or at least acknowledge its contradictions. Consider the 'pre-cognition' studies of US social psychologist Daryl Bem, which were completed with remarkable transparency (D. J. Bem J. Pers. Soc. Psychol. 100, 407-425; 2011). (The general consensus is that these studies did not establish the presence of extrasensory perception in college students, but the prevalence of overly flexible statistics; Bem defends the statistics as sound.) The work flouted well-supported ideas about physics and causality. It was akin to when physicists at CERN, Europe's particle-physics laboratory near Geneva, Switzerland, 'discovered' faster-than-light neutrinos, violating the special theory of relativity. Because the researchers required

We don't just want science to be reproducible. We want it to help us to make better sense of the world,"

Paul Smaldino is an assistant professor of cognitive and information sciences at the University of California, Merced. psmaldino@ ucmerced.edu

their results to be consistent with a broad theoretical framework, they probed deeper and discovered that their finding stemmed from a loose fibre-optic cable. To be clear, it's not the case that surprising claims are always wrong – but such claims must undergo extensive scrutiny.

If useful models produce better science, then what drives better models? Improved measurements. Consider the work of Tycho Brahe - a great astronomer of the sixteenth century, who nonetheless thought that the Sun orbited Earth. Yet his painstaking measurements of the positions of the planets allowed Johannes Kepler to determine that their orbits are elliptical. From this, Isaac Newton could formalize his theory of universal gravitation, which allowed modern researchers to ask countless questions about planetary motion, cosmology, ballistics, engineering and more.

If we can't reliably measure something, it's hard to build a theory about it. Quantities such as position, mass and time are relatively easy to measure, at least at some scales. Cognitive scientists can readily measure skin conductance, reaction times and word counts; this allows regularities and variation to be observed, and thus the construction of testable models. Other fields, including those I work in, have struggled with measurements. Psychologists attempt to measure emotions, identities and beliefs. Social scientists attempt to measure inequality, polarization and disinformation. Biomedical scientists attempt to measure treatment outcomes in small, heterogeneous populations.

I think that many sciences struggling with replication are those with the most pressing challenges in taking clear measurements. The trick lies not in merely finding a measurement that can be made precisely or described transparently, although these factors are important. Instead, scientists must find properties that can be reliably measured, inform theory and lend themselves to quantification in formal models.

Ideally, strong theories, formal models and measurements will interact in a virtuous cycle. Models allow us to study assumptions about the world and discover their consequences. The results can show what measurements are needed to test the assumptions, and those measurements can provide empirical patterns that invite explanations, which models can provide. And on and on.

We absolutely need better methods for hypothesis testing, and these are already being incorporated into how scientists are trained and how science is done.

So now it is time to focus on better practices for hypothesis generation. We need training programmes in model building and critique, plus consortia-building and funding programmes to invent and test measurements that make models tractable.

Better methods will help us get the right answers; models and measurements will ensure we ask the right questions.

## **News in brief**

### CLIMATE FUND ATTRACTS RECORD SUM FOR **DEVELOPING NATIONS**

Developed nations have together pledged US\$9.8 billion to replenish a United Nations fund that helps low-income countries to reduce their carbon emissions and adapt to the impacts of climate change.

At a conference on 24-25 October in Paris. 27 countries promised to contribute to the latest fund-raising round for the Green Climate Fund (GCF). The total value of these pledges exceeds the \$9.3 billion promised in the last round in 2014, despite the absence this time of the United States and Australia. Thirteen nations. including the United Kingdom, Germany and France, pledged at least double what they did five years ago, in domesticcurrency terms.

The fund was established in 2010 and has so far allocated \$5.2 billion to climate-change mitigation and adaptation

projects around the world.

The United States committed more money to the GCF than any other nation in 2014, but US President Donald Trump has since withdrawn \$2 billion of the \$3 billion that was promised, and has declined to contribute further to the fund. This left a substantial hole in the GCF's coffers, although European nations have largely made up the shortfall.

The fund remains open, and it is likely that more countries will make pledges in the coming months. Countries that have been stymied by domestic political processes could also increase the amount they have said they will give. More funding is expected from Belgium, for instance, where a parliamentary resolution to double its \$45-million contribution came too late to be reflected in its most recent pledge.



# INTERSTELLAR COMET CONTAINS

Astronomers have for the first time spotted signs of water in our Solar System that originated somewhere else. The alien water seems to be spraying off comet 2I/Borisov, which is flying towards the Sun on a journey from interstellar space, reported a team led by Adam McKay at NASA's Goddard Space Flight Center in Greenbelt, Maryland, on 28 October (A. McKay et al. Preprint at https://arxiv.org/ abs/1910.12785; 2019).

"There's water – that's cool, that's great," says Olivier Hainaut, an astronomer at the **European Southern Observatory** in Garching, Germany, Most comets contain a lot of water. he says – but confirming its presence in an interstellar comet is an important step towards understanding how water might travel between the stars.

Astronomers have been avidly tracking Borisov ever since its discovery on 30 August. It is only the second interstellar object ever spotted.

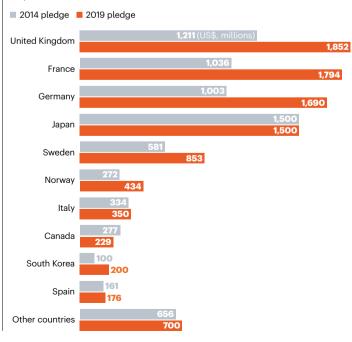
McKay and his colleagues used a 3.5-metre telescope at Apache Point Observatory in Sunspot, New Mexico, to probe the sunlight reflecting off Borisov. On 11 October, they spotted signs of oxygen in light coming from the comet. Although comets can produce oxygen in various ways, researchers say that the most likely source is water breaking apart into hydrogen and oxygen.



**Major climate** conference swaps venue amid protests

#### **CLIMATE CASH**

In the latest fund-raising session, 27 countries pledged US\$9.8 billion to the Green Climate Fund.



. TO R: SOURCE: GREEN CLIMATE FUND; D. JEWITT (UCLA)/ESA/NASA; ALBERTO VALDES/EPA-EFE/SHUTTERSTOCK; EZRA ACAYAN/GETTY; NEWSCOM/ALAMY



Sign up to get essential science news, opinion and analysis delivered to your inbox daily. Visit go.nature.com/newsletter



The United Nations' annual climate summit will decamp to a new continent, as a result of massive protests against economic inequality (pictured) that have rocked Chile for nearly two weeks.

On 30 October, Chilean President Sebastián Piñera cancelled plans to host the climate meeting, known as the 25th annual Conference of the Parties (COP25), which was due to start in December in Santiago, citing safety concerns. A day later, Spanish President Pedro Sánchez offered to host the summit in Madrid, a proposal the UN has accepted. The summit will still take place on 2-13 December.

Countries attending COP25 plan to work out the details of implementing the Paris climate agreement ahead of 2020, when they will update their climate pledges under the international pact. As many as 25,000 people are expected to attend the talks.

The cancellation was the latest in a series of obstacles for the climate conference. Chile had agreed last year to host the talks after Brazil backed out of holding the meeting.

# MEASLES ERASES IMMUNE 'MEMORY' FOR OTHER DISEASES

Measles infections in children can wipe out the immune system's memory of other illnesses. This can leave children vulnerable to other pathogens that they might have been protected from before their bout of measles.

The findings, published on 31 October in Science and Science Immunology, come at a time when measles cases are spiking. Globally, there were more measles infections in the first six months of 2019 than in any year since 2006, according to the World Health Organization.

The measles virus seems to destroy immune cells that 'remember' encounters with specific bacteria and viruses (V. N. Petrova et al. Sci. Immunol. 4, eaay6125; 2019). And results from a separate team indicate that measles can damage plasma cells in the bone marrow, cells that could otherwise produce pathogen-specific antibodies for decades (M. I. Mina et al. Science 366, 599-606; 2019).

The findings emphasize how the measles vaccine protects against more than just measles, says Velislava Petrova, an immunologist at the Wellcome Sanger Institute in Hinxton, UK, who led the Science Immunology study.





# SOUTH KOREA CLAMPS DOWN ON

South Korea's education ministry wants to stop academics from participating in conferences that have little academic value. The ministry announced on 17 October that it will require all universities to adopt measures to vet academics' travel to overseas conferences so as to "prevent researchers from engaging in poor academic activities".

The ministry's order comes after a report that it released in May, which found that 574 professors from 90 universities around the country had participated in conferences that it called "weak". It is thought that some researchers knowingly elect to pay the fees to attend such conferences, or to publish in low-quality journals - some of which are considered 'predatory' journals – because they are a quick and easy way to add a publication or presentation to their CVs, or to gain experience in presenting at international conferences.

Under the new policy, researchers will be required to fill out checklists before attending overseas conferences and then submit the lists to their universities, which will use them to screen the adequacy of the researchers' academic and research activities, the ministry told Nature in a statement.



The Kincade Fire has burnt a swathe through Sonoma County, California, since it began on 23 October.

# CALIFORNIA SCIENTISTS RACE TO ASSESS HEALTH RISKS OF WILDFIRE SMOKE

Bay Area study will track long-term effects of pollution on the heart, lungs and immune system.

#### By Amy Maxmen

s the skies above the San Francisco Bay Area in California filled with smoke in late October from wildfires ripping through nearby Sonoma County, Kari Nadeau and Mary Prunicki sprang into action.

The pair, scientists at Stanford University in the Bay Area, began calling in hundreds of people who had signed up to participate in their study of the long-term health effects of wildfire smoke. Previous research has linked air pollution from wildfires to surges in hospital visits for asthma and strokes. But it's not clear whether exposure to wildfire pollution creates

chronic health problems – something that Nadeau, director of Stanford's Sean N. Parker Centre for Allergy & Asthma Research, and Prunicki, a pollution biologist, hope to find out.

In early October, before the first large wildfires of the year sparked in northern California, their team assessed the circulatory, respiratory and immune systems of people enrolled in the study. The scientists began calling participants back to their lab on 28 October to undergo the same tests, which they will repeat in three months after the smoke has cleared. Nadeau and Prunicki have approval to continue assessments until 2037, and ultimately hope to enrol as many as 2,000 people - amassing a trove of data on how a person's

body responds to wildfire smoke over time.

Answers are sorely needed. Wildfires burned a record-breaking 760,000 hectares last year in California; almost 100 people died and hundreds of thousands of others breathed in sooty air for days. As Nature went to press, the massive Kincade Fire in northern California had burnt about 32,000 hectares, destroyed more than 370 structures and prompted evacuations and power outages (see 'Wildfires disrupt science'). And climate models predict that such blazes will grow larger and more frequent in the coming decades. The area burnt in California each year will increase by 77% by the end of the century if greenhouse-gas emissions continue to rise, according to the state's

#### **News in focus**



The area burnt by wildfires in California is projected to rise as climate change intensifies.

most recent climate-change assessment.

Lisa Miller, an immunologist at the University of California, Davis, says the Stanford study is one of the first to monitor wildfires' health effects in a diverse group of people over several years. By understanding who is most affected by wildfire and why, Miller says researchers can create evidence-based guidelines for mitigating risk. She is particularly worried that smoke exposure could damage children's developing lungs in ways that lead to chronic health problems.

"We have to be better prepared for these events," she says. "Last year was everyone's wake-up call that we need to be ready for the next big fire to happen."

The idea for the health study arose last year. as the largest and deadliest blaze in California's history – the Camp Fire – ravaged the northern part of the state. After the fire destroyed the city of Paradise in California's Central Valley and turned skies brown above the Bay Area, Nadeau and Prunicki realized their skills were needed.

The pair has long studied how air pollution in the central California city of Fresno alters immune cells and causes allergies and asthma. In April, they reported that 7- and 8-year old children living about 100 kilometres away from wildfires in 2015 were exposed to more pollutants than were those living near prescribed burns - small forest fires that are purposely set to reduce overall fire risk (M. Prunicki et al. Allergy 74, 1989-1991; 2019).

The researchers suspect that the difference is due to toxic chemicals released when wildfires burn synthetic materials in houses and cars. "Wildfire is like a giant slug of air pollution all at once," Prunicki says.

As the smoke from the Camp Fire hung over the Bay Area last year, she and Nadeau

scrambled to launch a small study tracking the effects of wildfire pollution on health. They collected blood and saliva from about 100 people, and asked them to return for assessments a few months later. "We didn't have time to collect tons of information, and it was sort of done in reverse," says Nadeau.

In February, they submitted a proposal for a larger study to an ethical review board. To fund the work, they set aside about US\$1 million from a grant they had received from the Parker Foundation in San Francisco.

The team is conducting the wildfire research in the Bay Area because the air quality is typically better than that of Fresno, says Nadeau. This should help her team to isolate the health effects of wildfire smoke from those caused by other environmental hazards.

The scientists also began an 80-person study last week to test whether air purifiers can limit any health effects from exposure to wildfire smoke. Half of the students in a college dormitory in Fresno have air-filter machines installed in their rooms, and the other half have a sham machine. The goal is to work out how much air filters help, and who needs them.

Michael Wara, an energy and climate policy analyst at Stanford, hopes to incorporate data from Nadeau and Prunicki's health studies into models on the costs and benefits of various policies to curb wildfire damage. "Fire is a climate-adaptation problem that California is confronting right now," he says, "and not in 2050, and not in 2100."

The researchers behind the northern California health study hope that its findings will help people around the world who are exposed to wildfire smoke. "This isn't just a problem for the [US] west," Nadeau says. "We need to know how to adapt better. Right now, people are left unaware."

### **Wildfires** disrupt science

Power cuts have added to uncertainty for researchers.

The blazes that have torn through California since late October have prompted evacuations and power outages that have disrupted research.

The University of California, Berkeley (UCB), and the neighbouring Lawrence Berkeley National Laboratory (LBNL) were among the institutions in northern California that closed on 26 October as a result of a planned blackout that followed the Kincade Fire, which broke out on 23 October near Santa Rosa.

This was the second outage in a month. The first, which occurred on 9-10 October when the Pacific Gas and Electric Company (PG&E) of San Francisco, California, cut power to reduce the risk of wildfires, caused the most chaos. One UCB lab moved freezers full of specimens to nearby facilities that still had power, while others stocked up on dry ice to keep their samples frozen.

But researchers say that things went more smoothly during the second blackout. That time around, university officials pre-emptively switched to a campus power plant before PG&E cut electricity to the area on 26 October. Despite the campus closure, researchers were still able to access facilities to check on their samples and experiments, but they had to scramble to relocate meetings.

A conference on the popular geneediting technique CRISPR, scheduled for 26 October, had to be moved off campus, says Jennifer Doudna, a biochemist at UCB. Organizers streamed the meeting online for those who couldn't squeeze into the smaller space. "How can we be living in a state with the fifth-largest economy in the world and having power outages like this?" Doudna asks.

She hopes the situation will push lawmakers and PG&E to bolster the grid to avoid such disruptions in the future. "I don't think this type of climate is going away," says Doudna. "We have to plan for it."

UCB resumed normal operations on 29 October, and LBNL reopened on 30 October. In Los Angeles, another blaze, the Getty fire, prompted the University of California, Los Angeles, to cancel classes for one day on 28 October.

By Jeff Tollefson

# MARK MACEWEN/NATURE PICTURE LIBRARY

# PRIMATE EMBRYOS GROWN IN THE LAB FOR LONGER THAN EVER BEFORE

The 20-day-old monkey embryos could reopen the debate about how long the human variety should be allowed to grow in a dish.



Two groups in China have grown embryos from cynomolgus monkeys for 20 days.

#### By David Cyranoski

They are the longest-lived primate embryos to thrive outside the body. The monkey embryos survived in a dish for 20 days, thanks to techniques developed by two groups working in China. The work sheds light on a crucial but little-understood phase of early development, and will probably reignite the debate about how long human embryos should be permitted to develop in the laboratory.

Researchers grow embryos to understand the earliest stages of development. In 2016, biologists in the United States grew human embryos in the lab for 13 days, but then stopped the experiments because of an internationally accepted rule not to allow growth beyond 14 days for ethical reasons. Because monkeys are a closely related species, their embryos are a window into early human development, but scientists have previously grown them for only nine days.

The two teams in China now report in Science<sup>1,2</sup> that lab-grown embryos from cynomolgus monkeys (Macaca fascicularis)

underwent several crucial processes. In one of these, gastrulation, the basic cell types that give rise to different organs begin to emerge, at around day 14.

"The best part is that there is a system to study gastrulation in vitro in a model very similar to the human," says Magdalena Zernicka-Goetz, a developmental biologist at the California Institute of Technology in Pasadena. "This is very exciting."

Although the studies show that early monkey development mirrors many aspects of the first two weeks of the human process, the teams report subtle differences between the two species. This suggests that monkey embryos might not be an adequate model for studying some advanced stages of human development. says Pierre Savatier, a stem-cell biologist at the Stem-cell and Brain Research Institute in Bron, France. He predicts that the papers will reinvigorate a push to extend the 14-day policy.

The ability to grow monkey embryos for longer than ever before could also boost research in another hot and controversial field - the generation of hybrid human-monkey embryos, known as chimaeras, with the goal of investigating how human cells differentiate into organs. This research has been held back because researchers haven't been able to grow monkey embryos for long enough to see how the injected human cells behave. Savatier says he will use the culture technique to grow monkey embryos that will be injected with human stem cells. "This culture system is hugely important for chimaera experiments," he says.

#### **Embryo bonanza**

Both teams grew monkey embryos on a gel matrix that supplied higher levels of oxygen than do cells in the womb. This culture technique was developed by Zernicka-Goetz's team, which was one of two groups<sup>3,4</sup> in the United States that grew human embryos for 13 days, in 2016.

In one of the latest two papers, a team led by Juan Carlos Izpisua Belmonte, a developmental biologist at the Salk Institute for Biological Studies in La Jolla, California, and Ji Weizhi at the Yunnan Key Laboratory of Primate Biomedical Research in Kunming, China, reports that 46 of 200 monkey embryos survived to 20 days. The authors of

#### **News in focus**

the other paper, led by Li Lei, a developmental biologist at the Institute of Zoology, Chinese Academy of Sciences, in Beijing, say they grew three embryos for that long.

The teams tracked the progress of the embryos, which were created using in vitro fertilization, to check whether they grew as they would have in the womb. They examined the timing and shape of structures in the embryos and the structures that support embryonic growth, the types of protein that are expressed by cells at different stages and the primordial germ cells that go on to become eggs or sperm. Then they compared these observations with what is known about development of this species from past experiments, in which embryos were removed from pregnant monkeys at different stages up to 17 days.

Both groups report that embryos in a dish develop in the same way as those in the womb. "It's ok to assume that the observations made are a representation of what happens in vivo," says Izpisua Belmonte.

The teams stopped their experiments on day 20, when the embryos turned dark and some cells detached - signs that the structures were collapsing. Li says it's not clear why that happened. He and Izpisua Belmonte say that culturing the cells in an extracellular matrix that better mimics the womb might help them to survive longer. Next, Ji hopes to grow embryos to the point when the primitive nervous system starts to form, around day 20.

#### **Sutble differences**

Savatier savs one difference between monkey and human embryos, described in the Ji and Ispizua Belmonte paper, is that the genes that are expressed in monkey cells that form the placenta are different from those in humans. But to study these processes in later stages in human embryos, regulators would need to lift the 14-day ban.

After the US teams grew human embryos to 13 days, some scientists and ethicists pushed for a revision of the policy, and a poll conducted in the United Kingdom in 2017 reported strong public support for extending the limit beyond 14 days. Savatier and others think the latest results showing the unique features of human embryonic development will strengthen arguments to change the policy.

Researchers are optimistic that the gel matrix could be used to grow human embryos to a more advanced stage if the rules change. li says that another group at his institute has developed a protocol specifically for human embryos that will soon be published. "This system could be suitable for human embryos to be cultured to 20 days," he says, "but we are not planning to do it."

- Niu, Y. et al. Science http://doi.org/ddn3 (2019).
- Ma, H. et al. Science http://doi.org/ddn4 (2019).
- Deglincerti, A. et al. Nature 533, 251-254 (2016).
- Shahbazi, M. N. et al. Nature Cell Biol. 18, 700-708 (2016).

# GENOMES TRACE ORIGINS OF ENSLAVED PEOPLE WHO DIED ON ISLAND

Former slaves left on St Helena were probably taken from west-central Africa, a genome study finds.

#### By Ewen Callaway

enomes from enslaved Africans who were freed and died on a remote Atlantic island in the mid-nineteenth century are offering clues about their origins in Africa. The findings come from the largest study of genome data obtained so far from remains of enslaved people and offer insights into the transatlantic slave trade, in which an estimated 12 million Africans were kidnapped and enslaved in North and South America and the Caribbean.

Researchers analysed DNA taken from 20 people from the British island territory of St Helena, whom the British Navy had liberated and brought there. The research, posted on the bioRxiv preprint server last

"By illustrating the history and the condition of a few, we are at the same time illustrating the condition of the many."

month, suggests that the people might have been captured in parts of west-central Africa, including present-day Angola and Gabon (M. Sandoval-Velasco et al. Preprint at bioRxiv http://doi.org/ddq2; 2019).

#### No island paradise

St Helena, which lies in the Atlantic Ocean nearly 2,000 kilometres west of Angola, occupies a unique chapter in the history of the transatlantic trade in people. After Britain outlawed the slave trade in 1807, its navy intercepted slave ships and sent an estimated 24,000 people to St Helena. They had been aboard ships heading largely to Brazil and Cuba between 1840 and the late 1860s.

Many of the people freed arrived in poor health and were housed in squalid conditions, and as many as 10,000 died. In 2006. construction work uncovered mass burials, and archaeologists unearthed the remains of 325 people – more than half under 18.

Unlike cemeteries in the Americas, which tend to hold multiple generations of people who had once been enslaved, nearly all of the people who died on St Helena were likely to have been born in Africa.

Shipping records – the main historical source on the African origins of people taken into captivity – tend to record only the ports from which slave ships set sail, but other records suggest that many of the people were captured farther inland.

In an attempt to better trace the Africans left on St Helena, a team led by palaeogenomicist Marcela Sandoval-Velasco and ancient-DNA researcher Hannes Schroeder, both at the University of Copenhagen, tested remains from 63 of the people for intact DNA. They sequenced partial genomes from 20.

Seventeen were male – backing up records indicating that, in its final decades, the transatlantic slave trade captured more men than women. Analysis of the genome data found that none of the people were closely related, nor  $did\,they\,belong\,to\,a\,single\,African\,population.$ 

Comparisons with genome data from thousands of modern Africans from dozens of populations suggest that the people from St Helena are most closely related to people living today in central Gabon and northern Angola. But the researchers caution that gaps in present-day genome data from potential homelands, such as the Democratic Republic of the Congo, make it difficult to say for certain where the people buried in St Helena were taken from. "Although it's very hard to exactly pinpoint their origins, I think what we see in our results is that they are not coming from a single population," says Sandoval-Velasco.

This insight suggests that the liberated Africans on St Helena lived in a challenging multicultural setting where they might not have understood the language and customs of others left on the island. "We hope that by illustrating the history and the condition of a few, we are at the same time illustrating the condition of the many, but it shouldn't stop there." Sandoval-Velasco savs.

Genome analysis shines a powerful light on people exploited in one of history's darkest chapters, says Rosa Fregel, a population geneticist at the University of La Laguna in the Canary Islands. "Usually it's just about numbers – how many people from each country. Here, we are talking about particular people and their origin," says Fregel. "Ancient DNA has the potential to tell their story."



Communities will receive 1.5% of the farm gate price, equivalent to US\$799,000 in 2019.

# INDIGENOUS COMMUNITIES WIN HISTORIC RIGHT TO ROOIBOS TEA PROFITS

Industry agrees to pay but contests research that San and Khoi used rooibos before European settlers.

#### By Linda Nordling

ore than a century after commercial farming began on their traditional lands, the San and Khoi peoples of southern Africa will share in the profits of the lucrative rooibos tea industry, the South African government announced on 1 November.

The announcement is the culmination of a decade-long negotiation between industry representatives and San and Khoi community groups. A review of the historical and ethnobotanical research literature published by the South African government in 2015 concluded that there is a "strong probability" that the first users of rooibos were the San people and that they - and the Khoi - should be compensated by industry (see go.nature.com/2jkviye).

Industry representatives have agreed to the payments, but say that they do not accept the government's interpretation of the research. They conducted their own literature review, which comes to a different conclusion.

By contrast, representatives of the communities and the government have welcomed the agreement - the largest of its kind between Indigenous peoples and industry.

"Rooibos is part and parcel of my upbringing," says Collin Louw, chairman of the San Council of South Africa. He says his ancestors used it to soothe skin rashes, among other things.

The agreement is also significant because it's the first such arrangement since the 2010

"These are sensitive issues. The concerns of centuries cannot just be resolved in a few years."

ratification of the Nagoya Protocol of the **United Nations Convention on Biological** Diversity. This is an international law that sets the rules for compensating communities if their knowledge of biodiversity is used by businesses or scientists.

Tim Hodges, a Canadian diplomat who co-chaired the Nagoya process, calls the rooibos agreement a historic achievement and a model for other countries and industries.

In particular, the announcement will be

closely read by researchers and funders involved in the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES), an effort to provide scientific advice on the world's epic loss of biodiversity.

According to the IPBES, traditional knowledge - defined by the World Intellectual Property Organization as knowledge that is handed down from generation to generation of biodiversity is key to discovering as-yet undescribed species. South Africa's decision suggests that brokering agreements to access this knowledge will take time.

"These are sensitive issues," says IPBES member Unai Pascual, an environmental economist at the Basque Centre for Climate Change in Bilbao, Spain. "The concerns of centuries cannot just be resolved in a few years."

#### Counting the cost

Under the rooibos agreement, the San and Khoi communities will receive 1.5% of the 'farm gate price' - the price that agribusinesses pay for unprocessed rooibos (Aspalathus linearis), which is endemic to the Cederberg region north of Cape Town.

For 2019, the government considers that the compensation will amount to 12 million rand. (US\$799,000). The communities will split the proceeds fifty-fifty. A third group small-scale non-white rooibos farmers in the region who were disadvantaged under apartheid – will share in the Khoi portion.

The San communities are among the world's oldest, and are understood to have been in Southern Africa for some 100,000 years. The Khoi arrived more recently, about 2,000 years ago. European settlers attacked these communities and occupied their lands starting in the mid-1600s, and today both San and Khoi peoples are scattered throughout southern Africa.

Commercial rooibos farming, which is now worth an estimated 500 million rand a year, began on these lands in the early 1900s, industry representatives say. In 2010, the San Council of South Africa approached the government with a claim under South Africa's biodiversity law, asking for compensation for its peoples' traditional knowledge of the plant and for the use of San imagery in rooibos packaging and marketing.

#### A question of origins

In spite of the agreement, the precise origins of rooibos tea remain contentious. Representatives of the San and Khoi say that their ancestors shared knowledge of the plant with colonial settlers. The literature also points to its early uses as a health tea and as a diuretic.

Overall, there is a lack of research literature in this field, but what there is suggests that rooibos as a beverage did originate with the groups' ancestors, according to the 2015 study commissioned by the South African government.

However, a separate 2017 report commissioned by the industry-led South African Rooibos Council (SARC) says that there is no conclusive recorded evidence that the original inhabitants of the Cederberg region used rooibos to brew tea, or that they taught colonial-era settlers about it (see go.nature. com/2ndv2zg).

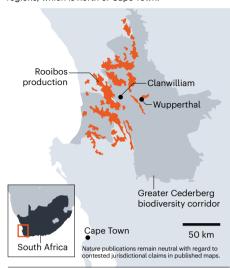
#### **Alternative interpretations**

This disagreement caused deadlock as each side stuck to its interpretation of the research. SARC chairman Martin Bergh, who is also managing director of one of the largest rooibos agribusinesses, Rooibos Ltd, based in Clanwilliam, South Africa, says the industry still does not accept that Indigenous communities used rooibos as tea. But he does agree that the San probably knew about the rooibos plant before anybody else.

However, the government accepted that the communities deserve to be compensated because rooibos is endemic to where it is now grown (see 'The rooibos belt'), and the San and Khoi lived there for centuries before the settlers. The 2015 study also found no evidence casting doubt on the communities' argument that their ancestors used rooibos as a beverage.

#### THE ROOIBOS BELT

Rooibos is endemic to Cederberg, one of South Africa's most biodiverse regions, which is north of Cape Town.



All sides have pledged to revisit the agreement in a year's time, because other questions also remain. For example, Rachel Wynberg, who researches the commercialization of biodiversity at the University of Cape Town, questions how the funds will reach San and Khoi individuals, many of whom are not well-connected with Indigenous leadership structures such as the San Council. She also says that establishing which community contributed how much to the plant's uses is "fraught with difficulties", and is likely to lead to arguments - especially where communities have been oppressed and marginalized.

And Barend Salomo, who manages a cooperative of small-scale farmers in Wupperthal who will also benefit, says that the money for his community will not stretch far. He hopes that the agreement can be tweaked to provide greater dividends in future years. "We don't want to kill the industry, but this is not fair,"

But Willie Nel, a large-scale rooibos farmer based outside Clanwilliam, says that if farmers are unable to recoup the cost of the levy by charging higher prices, they will lay off farm labourers, who, he says, are mostly members of the third group that the agreement is supposed to help. Nel also worries that mounting demands for restitution might lead to the industry pricing itself out of the market. "We think rooibos is special but there are so many teas from all over the world. And we have to compete with all of them."

But as more than 80% of rooibos tea is sold in Europe, Japan and North America, environmental economist Pascual doesn't think these sales are likely to be affected by a small increase in price. "That's not how economics works," he says.

### nature masterclasses

#### Online Course in Scientific Writing and Publishing

Delivered by Nature Research journal editors, researchers gain an unparalleled insight into how to publish.



Try a free sample of the course at masterclasses.nature.com



Bite-size design for busy researchers • Subscribe as a lab or institution

W masterclasses.nature.com

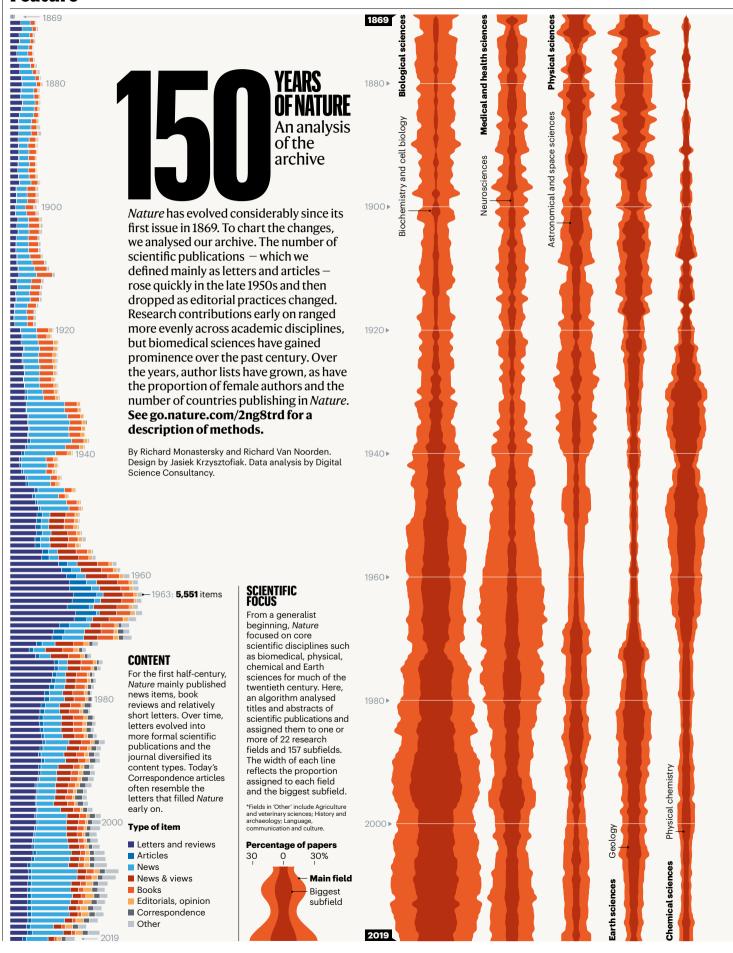


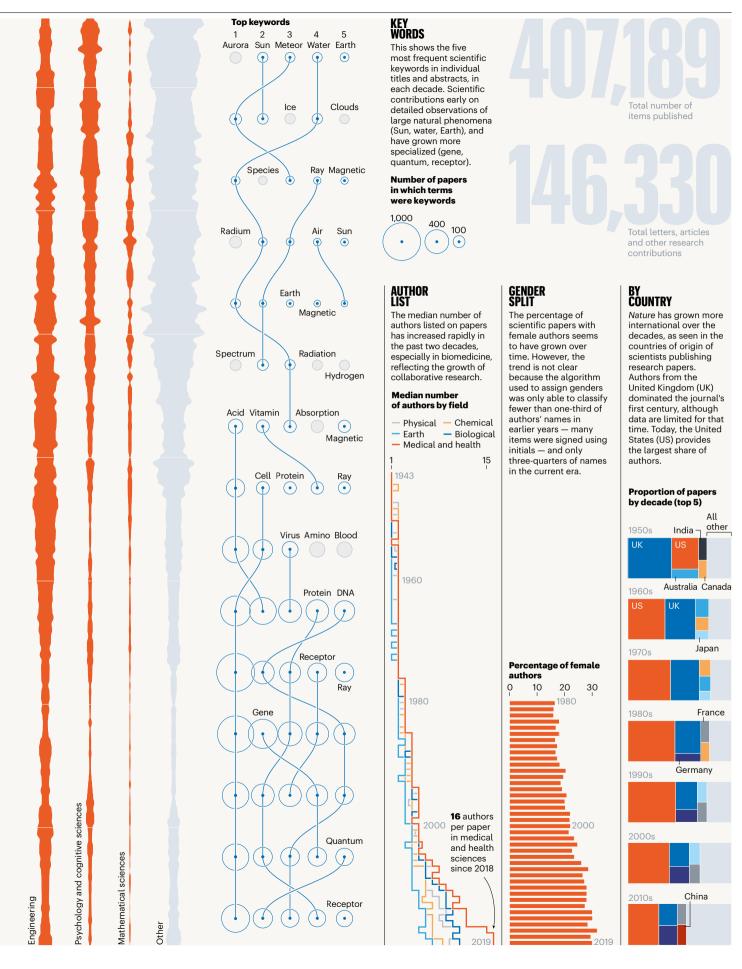
in Follow us on LinkedIn



Skills and Careers Forum for Researchers

#### **Feature**





## Red end. Violet end. hydrogen hydrogen maanesium hydrogen

Fig. 1.—Showing the solar spectrum, with the principal Fraunhofer lines, and above it the bright-line spectrum of a prominence containing magnesium, sodium, and iron vapour at its base.

Norman Lockyer's figure of a solar spectrum, published in the first issue of Nature.

## The evolution of scientific illustration

#### The interplay of image-making, research and visual technologies over the past 150 years. By Geoffrey Belknap

cience is a fundamentally visual endeavour. It pivots on the material whether that is an atom, a gene, a crystal, a whale or a distant galaxy. Its aim is elucidation. Thus, communicating research has always been predicated on combining image and text to share discoveries. ideas and observations.

When it came on the market in November 1869. Nature stated its commitment to the visual with a beautifully drawn masthead showing Earth emerging from clouds. (The artist might have been engraver James Davis Cooper, who illustrated Charles Darwin's 1872 book Expression of Emotions in Man and Animals.) Under the masthead were the words 'A Weekly Illustrated Journal of Science'. The banner, if not the subtitle, remained on Nature's front page until just after the Second World War.

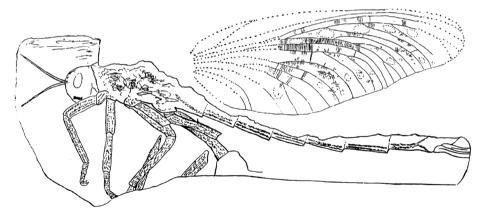
Over the years, Nature adapted through its succession of editors, with, in recent decades, 'sister' journals carving out their own space in increasingly specialized scientific disciplines. Images remained central throughout. For instance, in 1896, Nature published physicist Wilhelm Röntgen's first X-ray plates<sup>1</sup>; in the 1920s, maps to debate Alfred Wegener's theory of continental drift<sup>2</sup>; and in 1968, the graphs that described astrophysicist Jocelyn

Bell Burnell's discovery of pulsars<sup>3</sup>.

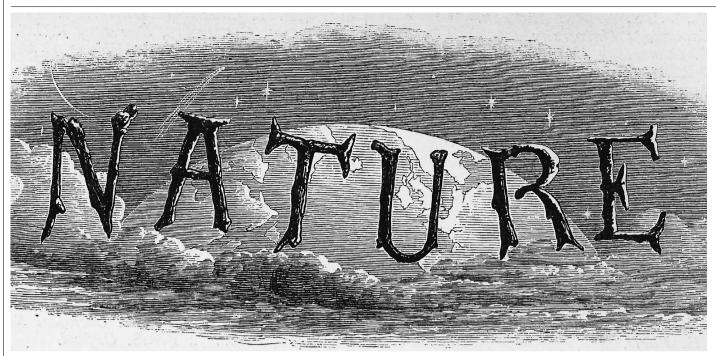
In some ways, the role of images in science publishing hasn't changed much over the past 150 years. Much scientific evidence takes the form of visualizations: illustrations, graphics and latterly photographs. What has shifted. inevitably, are the tools. Initially, Nature and other science journals featured monotone printed engravings. Now, its visual landscape is digitized, often mobile, presented in vivid colour and vastly expanded to reflect changes in technological capacity and science itself,

as the cover of this anniversary issue attests.

The late nineteenth century saw intense flux in scientific disciplines; boundaries between them were more porous. Images of discoveries sat cheek by jowl on journal pages with diatoms and archaeological artefacts (such as 800-450 BC stone tools found in Scotland by archaeologist Robert Munro<sup>4</sup>). Whereas many images are now used to interpret or visualize data, these early examples were mainly representations of scientific data - a photograph of an eclipse, say, or a drawing



A drawing of 'Titanophasma Fayoli Brongniart', published in Science in 1883.

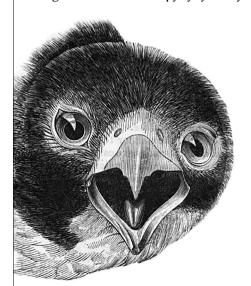


The illustration on Nature's inaugural cover, in 1869.

of a geological formation.

In Nature's inaugural issue, such data-led, representational images had an important role. The journal's first editor, astronomer Norman Lockyer, had co-discovered helium in 1868 using electromagnetic solar spectroscopy. He illustrated his article 'The Recent Total Eclipse of the Sun' with two photographic images: a solar spectrum and an engraving derived from a photograph of a solar eclipse.

These are not photographs as we would understand them. Before the 1890s, most images reproduced for journals were wood engravings, which were inked and set for printing alongside the typeset. To make them, the engraver would either copy by eye or lay



Drawing in the Magazine of Natural History.

a photograph directly onto the woodblock while carving.

#### The art of precision

Precision was centrally important. If a line on Lockyer's spectrum was in the wrong place, it might suggest that the Sun is composed of calcium rather than hydrogen. So, to ensure accuracy, skill and collaboration were necessary. The era's illustrators and engravers were often scientists themselves, or worked closely with researchers. Illustrators might even copy the image directly onto the block or plate ready for the engraver to do their work.

In their book Objectivity (2007), science historians Lorraine Daston and Peter Galison describe such collaborative processes of image-making as "four-eyed sight". Authors and image-makers worked together to shape and construct an observationally reliable image<sup>5</sup>. Similar collaborations were common throughout the century. For the first issue of his Magazine of Natural History, for instance, botanist John Claudius Loudon had an engraver copy the prints from John James Audubon's Birds of America (1827–1838). These worked as field guides for readers, even as the magazine became a forum for debating new findings with an expert community.

By the time Nature appeared, the model of a journal targeting a professional scientific community was emerging. Researchers might be 'amateur' naturalists who collected and described species, such as the botanist Alfred  $William\,Bennett\,and\,cryptogam ist\,Miles\,Joseph$ Berkeley, who sent images to Nature depicting the cause of 'rust' on wheat and barberry plants<sup>6</sup>. Or they might belong to the nascent class of university-based laboratory scientists

such as physicist Peter Guthrie Tait, who sent in a sketch visualizing his apparatus for measuring the wavelength of monochromatic light<sup>7</sup>. Specialist journals such as the Journal of Physiology, launched in 1878, reflected emergent disciplines while also making space for both amateurs and professionals.

Compelling imagery was becoming a competitive factor in this burgeoning marketplace

#### "Compelling imagery was becoming a competitive factor in this burgeoning marketplace of ideas."

of ideas. The non-scientific Illustrated London News, launched in 1842, had established a precedent for capturing large readerships through abundant visuals. As I described in my book From a Photograph (2016), Nature responded to this pressure to some degree<sup>8</sup>. Yet as historian Melinda Baldwin points out in Making Nature (2015), it wasn't until 1890 that the journal first made a profit<sup>9</sup>. The cost of images had a big effect on the bottom line. The geologist Edward Charlesworth, who took over ownership of the Magazine of Natural History from Loudon, had to slim issues down to squeeze in more images on each page10.

Nature, with its all-encompassing one-word title, also catalysed direct competition, such as astronomer Richard Anthony Proctor's Knowledge (subtitled 'An Illustrated Magazine of Science') and the US illustrated weekly Science. In 1883, one of the latter's early editors, the entomologist and palaeontologist Samuel Hubbard Scudder, published a description of a giant

fossil stick insect discovered in coal deposits in France by another entomologist. Charles Brongniart. Its accompanying engraving (see page 25) effectively stitches together two pieces of observational data – the body and wings of the insect, separated in the coal bed.

Such simple line engravings had become a staple. The same year in Nature, Canadian botanist Grant Allen published a series depicting the shapes of leaves, arguing that their shapes reflect levels of competition with other plants for access to energy sources.

Between the 1880s and 1900, the old collaborations gave way to technological interlocutors: photographers. Science journals viewed photography as a way of seeing that enabled "mechanical objectivity". There was greater trust in the power of ground lenses and silver halides to capture the world in a way that the eye cannot. As with all visual technologies, however, it needed selection, organization and interpretation for data to be rendered into a comprehensible image.

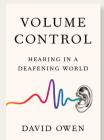
The work of French physiologist Étienne-Jules Marey is an iconic example. Following Eadweard Muybridge, who had captured animallocomotion through 'instantaneous photography' in the 1870s, Marey developed his own method: chronophotography. In an 1882 issue of Nature, he described his 'photographic gun', which used a rotating photographic plate to take sequential images of a flying bird11, helping to pave the way to understanding powered flight. Meanwhile, the Carte du Ciel project at the Paris Observatory, which ran from 1887 to 1950, led to the creation of 22,000 glass-plate negatives of stars from more than 20 observatories<sup>12</sup>.

Throughout the first half of the twentieth century, photography became crucial to science. Kathleen Lonsdale pioneered the form of crystallography in which X-rays are directed at a sample to measure diffraction and determine its atomic and molecular structure. Lonsdale published her 1928 findings on the benzene ring in the *Proceedings of the Royal Society*<sup>13</sup>. Her X-ray diffraction photographs – such as the 1941 series of eight, on diamonds - regularly appeared in *Nature*'s pages<sup>14</sup>.

The technique became crucial to the explosive discovery15 of DNA's structure by molecular biologists James Watson and Francis Crick, published in Nature in 1953. The key piece of evidence, 'Photograph 51', showed the diffraction pattern of DNA and was taken, under the supervision of crystallographer Rosalind Franklin, by then-graduate student Raymond Gosling<sup>16</sup>.

Photography was also used to disprove one of the biggest twentieth-century scientific hoaxes. In 1912, amateur archaeologist Charles Dawson claimed to have discovered the missing link between humans and apes in what looked like an early human skull found in Piltdown, Sussex. In 1913, the anatomist David

#### **Books in brief**



#### Volume Control

David Owen Riverhead (2019)

"For a deaf child, having hearing parents can be a serious handicap," notes New Yorker staff writer David Owen in this sensitive study of hearing. (He is personally involved, as someone with tinnitus who saw his grandmother struggle with deafness.) Meshing the science with individual auditory experiences, Owen discusses hearing aids, cochlear implants, genetically deafened mice, sign language, Thomas Edison and noise levels in US cities and towns — all in absorbing, anecdotal detail, although regrettably with no diagrams.



#### **Reality Ahead of Schedule**

Joel Levy Smithsonian (2019)

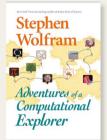
This picture-packed volume by science journalist Joel Levy tours scientific advances sparked by ideas in science fiction. The title comes from a definition of sci-fi by Syd Mead, an industrial designer behind the look of futuristic movies such as Blade Runner (1982). But how prescient is sci-fi? Levy shows how H. G. Wells's 1903 story 'The Land Ironclads' inspired Winston Churchill to promote the development of the military tank in 1915. But Wells did not envisage its key technical idea: caterpillar tracks, for added grip.



#### Jet Stream

Tim Woollings Oxford University Press (2019)

The jet stream — strong high-altitude air currents — was discovered in the 1920s. In this analysis of its complex impact on weather, physicist Tim Woollings relates how in 1944, the Japanese used the jet stream to launch trans-Pacific incendiary balloons. By strange chance, one hit the US plant that provided plutonium for the bomb that devastated Nagasaki in 1945. Today, argues Woollings, the jet stream is "very likely" to be threatened by another product of human activity: rising carbon dioxide emissions.



#### **Adventures of a Computational Explorer**

Stephen Wolfram Wolfram Media (2019)

Computer scientist and businessman Stephen Wolfram, designer of the technical-computing system Mathematica, proffers good stories in this collection of autobiographical essays. In 'Something I learned in kindergarten', he recalls himself as a six-year-old spotting a bite taken out of the Sun: a solar eclipse, something unknown to the other children. In 'My life in technology', he recalls rejecting the Latin word mathematica, learnt at school, as too long and ponderous. Silicon Valley luminary Steve Jobs convinced him otherwise.



#### Lightspeed

John C. H. Spence Oxford University Press (2019)

Starting with Albert Einstein, scientific consensus holds that the speed of light is a universal constant. So writes physicist John Spence in his history of attempts to measure the speed of light. Spence considers the implications of its constancy for modern physics and technology. For instance, the aether — a theoretical spacefilling medium rejected in Einstein's relativity — is still "anything but empty". Despite its appealing vignettes of great physicists, this is a challenging read. Andrew Robinson



Nature cover illustration from August 2019.

Waterson published a *Nature* article including three drawings taken from X-rays, revealing that the mandible of the Piltdown 'skull' was almost identical to a chimpanzee's<sup>17</sup>. By the 1950s, 'Piltdown Man' had been debunked. *Nature* had become a site for exposing bad science through visual evidence.

By the mid-twentieth century, the long-term boom in science journals (there were already 10,000 by 1900)<sup>18</sup> was unabated, keeping pace with the growth in academic science and the proliferation of fields. Photographic technologies remained central, but visual content was diverse and included graphs and early digital images. And starting in the 1970s, technological imaging innovations allowed science to see further and deeper. The cryo-electron microscope, first announced in *Nature*<sup>19</sup>, allowed electron microscopy to be applied to

organisms by freezing and suspending them in an aqueous solution. (In 2017, its inventors – James Dubochet, Joachim Frank and Richard Henderson – won the chemistry Nobel prize for their work on the structure of viruses.)

The invention of the charge-coupled device in 1969 meant that images could be captured on a silicon chip: photography had entered the digital realm. Digital-imaging sensors in telescopes have had a vast impact on astronomy. In 2018, Michael Koss and colleagues verified

"Macro to micro, imaging today is exquisitely precise and often beautiful."

the theory that black holes merge through the use of visual data from the Sloan Digital Sky Survey<sup>20</sup>. And in 2019, the first image of a black hole was released, created using the Event Horizon Telescope. *Nature* was key to communicating these new technologies and providing a platform for debating them.

Now, digital imaging is reaching further, with techniques such as hybrid multiplexed sculpted light microscopy under development to measure neuroactivity<sup>21</sup>, and NASA's grand database of multi-wavelength images of the galaxy (see go.nature.com/2bdrua5). There are extraordinary shots of nebulae snapped by the Hubble Space Telescope, 'extreme zoom' images of single atoms, and the burgeoning field of data visualization – graphical representations of data.

Macro to micro, imaging today is exquisitely precise and often beautiful, able to capture worlds and structures far beyond the scope of human vision. The 'Drawn Together' cover image for the 8 August 2019 issue of *Nature* is a case in point. Crafted by illustrator Inna-Marie Strazhnik, it provides a visualization of work by bioengineer Tyler Ross and colleagues, who used light-activated motor proteins to move microtubules into networked structures<sup>22</sup>. Strazhnik translated the models, images and graphs in the paper into a dynamic, almost three-dimensional image.

The visual continues to work as a foundation for making sense of data. The tools, as we have seen, have radically changed. The power of images has not.

**Geoffrey Belknap** is a historian of science, photography and visual culture. He is head curator of the National Science and Media Museum in Bradford, UK which is a part of the Science Museum Group.

- 1. Röntgen, W. C. Nature **53**, 274–276 (1896).
- 2. Wright, W. B. Nature 111, 30-31 (1923).
- Hewish, A., Bell, S. J., Pilkington, J. D. H., Scott, P. F. & Collins, R. A. Nature 217, 709–713 (1968).
- 4. A Scottish Crannog. *Nature* **22**, 13–16 (1880).
- Daston, L. & Galison, P. Objectivity (Zone Books, 2007).
   Bennett, A. W. & Berkeley, M. J. Nature 2, 318–319 (1870).
- 7. Tait, P. G. Nature 22, 360-361 (1880).
- Belknap, G. From a Photograph. Authenticity, Science the Periodical Press, 1870–1890 (Bloomsbury Academic, 2016).
- Baldwin, M. Making Nature: The History of a Scientific Journal (Univ. of Chicago Press, 2015).
- 10. Loudon, J. C. Mag. Nat. Hist. 9, iv (1836).
- Instantaneous Photography of Birds in Flight. Nature 26, 84–86 (1882).
- The Photographic Chart of the Heavens. Nature 38, 38 (1888).
- Lonsdale, K., & Whiddington, R. Proc. R. Soc. Lond. Ser. Contain. Pap. Math. Phys. Character 123, 494–515 (1929).
- 14. Lonsdale, K. & Smith, H. Nature 148, 112-113 (1941).
- 15. Watson, J. D. & Crick, F. H. C. *Nature* **171**, 737–738 (1953).
- 16. Franklin, R. E. & Gosling, R. G. Nature 171, 740-741 (1953).
- 17. Waterston, D. Nature **92**, 319–319 (1913).
- Shuttleworth, S. & Charnley, B. Notes Rec. R. Soc. Lond. 70, 297–304 (2016).
- Dubochet, A. M., Lepault, J. & McDowall, A. W. Nature 308, 32–36 (1984).
- 20. Koss, M. J. et al. Nature **563**, 214–216 (2018).
- 21. Weisenburger, S. et al. Cell 177, 1050-1066 (2019).
- 22. Ross, T. D. et al. Nature **572**, 224–229 (2019).

## **Comment**



## Science must move with the times

Philip Ball

Research cannot fulfilits social contract and reach new horizons by advancing on the same footing into the future, argues Philip Ball in the last essay of a series on how the past 150 years have shaped today's science system.

n 1866, three years before the first issue of *Nature* was published, a transatlantic telegraph cable established light-speed communication between Great Britain and North America. The triumph won William Thomson (later Lord Kelvin) a knighthood for the scientific advice he had given to the project. Yet Thomson had also advised on a disastrous earlier attempt in 1858 that barely worked from the outset and deteriorated within weeks.

It was partly in response to that costly debacle that the Cavendish Laboratory was established at the University of Cambridge, UK, in the early 1870s, to provide the nation's future engineers with a better grounding in physics. The

first director was James Clerk Maxwell, whose electromagnetic theory of the mid-1860s led to the discovery of radio waves in 1887 – which soon enabled 'wireless' telecommunication and rendered the telegraph obsolete.

In such ways, the distinctly Western and specifically British world into which Nature was launched regarded fundamental scientific research as the engine of socially transformative industrial innovation. Emanating from London, Norman Lockyer's journal showcased those developments from the perspective of a British Empire that grew to encompass about one-fifth of the world's population by the century's end. The benefits of research laboratories and the systematic institutionalization of science, in both academia and industry, were beyond doubt for Nature's target audience.

Eight decades later, this model motivated Vannevar Bush's 1945 report to US president Franklin D. Roosevelt. Science - The Endless Frontier made the case for governmental support of basic science research to promote national security, public health and welfare. It led to the establishment of the US National Science Foundation, and it appealed to the

#### Comment

optimistic and simplistic vision of science as a quest that, motivated by curiosity and guaranteed freedom of enquiry, would serve the interests of the nation and of humankind.

Science - whether it is Maxwell's electromagnetism, the Manhattan Project that inspired Bush, or the Human Genome Project - has indeed been so socially transformative that its intellectual and technological machinery has gained seemingly irresistible momentum. Is this not how progress is made, and is that not, on balance, a good thing?

Even to ask the question invites familiar and polarized arguments. Some commentators question the wisdom of unfettered scientific development, pointing to the problems of climate change and environmental despoliation, nuclear weapons and antibiotic resistance, along with the ambivalent influence of artificial intelligence and robotics, information technologies and genetic engineering. Others point out that quality-of-life indicators - lifespan and infant mortality, say – have improved steadily (if unevenly, geographically and temporally) during the era of modern science that roughly coincides with the span of Nature's existence.

But Manichean views and tropes of 'dual use' miss the point. Some of the key questions that confront science today are about whether its methods, practices and ethos, pursued with very little real change since Maxwell's day, are fit for purpose in the light of the challenges - conceptual and practical – we now face. Can science continue to fulfilits social contract and to reach new horizons by advancing on the same footing into the future? Or does something need to shift?

#### **Looking out**

Let's consider where we stand. The convention of the past century or so has tended to place the frontiers of knowledge at the scales of the very large and very small. Today we might be inclined to add the very complex - which typically pertains to the intermediate scales of direct human experience.

It's now clear that challenges at the two extreme scales – fundamental particles and cosmology – are related. As the island of knowledge grows, so does the perimeter of the horizon where knowledge ends, says Marcelo Gleiser, a particle cosmologist at Dartmouth College in Hanover, New Hampshire. "The more we know, the more exposed we are to our ignorance, and the more we know to ask", he writes1.

We have known for only several decades that dark matter outweighs all visible matter by a factor of five, yet we are no closer to knowing what it consists of. And scarcely two decades have passed since the mysterious entity dubbed dark energy, which causes the Universe's expansion to accelerate, has been recognized to comprise more than two-thirds of the total cosmic energy density. Never before has our knowledge of the Universe seemed so deficient.

Plugging these gaps at the largest scales will

depend on elucidating the physical world at the smallest. Here the prospects are currently dimenough to cause desperation and even rancour. The world's largest particle accelerator, the Large Hadron Collider at CERN near Geneva. Switzerland, has so far failed to offer any hint of how to proceed beyond known physics. Elegant ideas look moribund in the face of an ugly lack of facts. In the meantime, models are being forced towards ideas, such as the multitude of universes now permitted by the inflationary model of the Big Bang, that seem to some critics to abandon the empirical basis of science itself.

Yet even as our view of the Universe becomes increasingly perplexing, it is being fleshed out as never before. In the 1860s, it was almost casually assumed that life would be common on other worlds. H. G. Wells's 1897 novel The War of the Worlds (informed by his reading of Nature) seemed all the more chilling because of the widespread belief – which persisted for another half-century - that there was indeed life on Mars. Seasonal changes of surface colour were interpreted as vegetation growth, and striations described by astronomer Giovanni Schiaparelli were notoriously ascribed by others to artificial waterways.

But the barren, sterile Martian landscape that the Viking landers revealed in 1976 confirmed a growing sense – stoked by the Apollo Moon landings and reflected in physicist Enrico Fermi's famous question about the apparent absence of alien visitations – that we are a lonely outpost in a bleak, lifeless cosmos. Well, no longer. Since the first discovery of an extrasolar planet orbiting a Sun-like star was reported in this journal<sup>2</sup> in 1995, around 4,000 sightings of such planets have now accumulated (and a 2019 Nobel prize).

It seems that planetary systems are the norm for other stars, and Earth-like planets far from

#### "It's still unclear when or whether we can exclude ourselves from the scientific frame."

uncommon. Already we know a little about the atmospheres of some of these worlds. With the launch of NASA's Transiting Exoplanet Survey Satellite last year, and the James Webb Space Telescope scheduled to launch in 2021, we will soon know much more. Researchers are now speaking plausibly about deducing within a lifetime if there is life elsewhere.

Where does all this leave us? The cosmological perspective could seem to perpetuate the sense of an unfolding Copernican revolution in making humankind even more peripheral. Not just an insignificant dot in a vast Universe, we're possibly an insignificant universe in a potentially infinite multiverse. It's hard to imagine a demotion more extreme.

There is another view that is anything

but Copernican. Here, habitable worlds are ubiquitous and we remain uncomfortably. almost absurdly, at the centre of things. In the inflationary multiverse, our presence is the explanation for the fundamental constants of nature. They might have different values in other universes, but the conditions necessary for our existence guarantee that we will see the ones we do.

The foundations of quantum mechanics (a topic once disreputable that now verges on fashionable) muddy the picture too. The 'many worlds' interpretation is more popular today than when US physicist Hugh Everett proposed it in the 1950s. It multiplies universes (in a manner distinct from the cosmological multiverse) and it multiplies each of 'us' beyond measure. Meanwhile, US theoretical physicist John Wheeler's 'participatory universe' and new interpretations such as QBism3 insist that quantum theory requires the observer's presence rather than being the abstract and objective framework that science usually supplies.

These ideas remain speculative. But they challenge the Newtonian promise of an impersonal mechanics.

#### **Looking in**

In other words, it's still unclear when or whether we can exclude ourselves from the scientific frame. This would have been no surprise to Maxwell. His conception of physical reality was predicated (no less than was Newton's) on a religious position that awarded humanity a special place.

This, of course, is where Charles Darwin also enters the frame. His ideas on evolution by natural selection, published in On the Origin of Species (1859) were still causing shock waves when Nature was founded. Two years after that, he delivered the final bombshell in The Descent of Man (1871). The significance of his ideas was not as an explosive charge placed underneath the church but as the opening salvo to a century and a half of debate about what it means to be human. If there was a struggle, it was not about which book to consult but about who had the most decisive authority. Within science, first evolutionary theory, then psychoanalysis, and now genetics and neuroscience, have all staked their claims.

On Nature's centenary, you might have placed your bets with the latter disciplines. Half a century later, it is less clear that they can offer the last word. Powerful new techniques applied to rapidly growing data sets, such as genomewide association studies4, have disclosed a clear and sometimes strong genetic component to almost every human behavioural trait we choose to study, as well as influencing health and disease. But a mechanistic understanding of genetic effects often remains remote. And for traits in which many - perhaps even several thousand – genes are implicated, it is not even clear if this is the right level at which to ascribe





The first six primary mirror segments for the James Webb Space Telescope.

causes for what we can see and measure.

The emerging picture of development and tissue function at the level of single-cell transcription (and perhaps soon of translation) adds a new layer of complexity<sup>5</sup>. Apparently identical cells in the same tissue can show a wide range of dynamic states of gene expression. It might be that the genome tells us no more about how an organism builds and sustains itself than a dictionary does about how a story unfolds. New methods, rather than finally answering old questions, could merely beggar them, shifting the goalposts entirely as genomics itself has done for notions of race.

Neuroscience, like genetics, has been restricted in the questions it can ask by the data it can gather. Functional magnetic resonance imaging remains a blunt tool, showing where things are happening in the brain (at rather coarse-grained resolution), but not what transpires. The idea that the human brain might be understood by exhaustive documentation and perhaps simulation of neuronal connections and firing patterns was challenged as soon as it was mooted (by the ill-fated European Human Brain Project<sup>6</sup>).

Here we arrive at one stretch of the 'complexity' frontier. If history is any guide, we should expect that understanding these complex systems will not emerge by drawing analogies with the latest cutting-edge technologies. Just as the brain is not (as was thought in the early nineteenth century) a battery, neither is it a computer; nor is the genome a digital list of parts. And more data, although extremely valuable as a resource, will not help us without new ideas. These are in short supply. As neurobiologist and historian Matthew Cobb at the University of Manchester, UK, writes, "no major conceptual innovation has been made in our

overall understanding of how the brain works for over half a century"7.

It's no surprise, then, that the 'hard problem' of consciousness is barely articulated, let alone understood. We are still at the stage where serious thinkers on the topic embrace the gamut of positions, from regarding it as an illusion to considering it the only valid starting point for a theory of human experience. That latter view harks back to how US psychologist William James ignored "the traditional antithesis between reality and appearance", as *Nature* put it in 1915 (ref. 8). As for claims that neuroscience has banished free will (for example, because decisions can be predicted from brain scans in advance of their conscious manifestation). saying that "your brain decides before you do" merely returns us to British philosopher Gilbert Ryle's famous regression of mental homunculi9.

#### **New views**

Among the ways in which science has changed over the past century and a half, three loom large. First, it is no longer driven by lone figures labouring in their laboratories, but has become a team effort that spans labs, departments, disciplines, institutions and continents. Second, it often relies now on data sets so vast that human brains cannot hope to hold or parse them all. Third, it increasingly confronts issues of global reach and even existential urgency - from climate heating and the need for a carbon-neutral economy, to epidemics and water security.

Yet these changing demands are not reflected in incentives, funding mechanisms, awards or popular narratives. Systemic biases – for example, in barriers to the entry and advancement of women and people from minorities, or in the demographic coverage of medical databases, or the prejudices that algorithms inherit from their makers - remain entrenched. Even science's internationalism is threatened by current political trends. To regard what biologist Thomas Henry Huxley in Nature's first issue called the "progress of Science" as an inexorable, triumphant forward march, today seems dangerously complacent.

It is time to ask whether such problems are not imperfections of the system but consequences of it. Science might be hindered by channelling its practitioners into a single mode of thinking. There is hubris in the assumption that the traditions, conventions, training, disciplinary boundaries, methods, responsibilities and social contract that crystallized in the nineteenth century from a highly restricted demographic must still be the best way of working. To say as much is not to submit to some trendy caricature of postmodernism. Rather, it is to acknowledge that there are assumptions embedded, often invisibly, in the way we develop models, deploy metaphors, apportion priorities, recognize and reward achievement, and recruit participants that must be questioned.

The canonical scientific article, with its unified and passive voice, its closed and self-contained narrative, its seductively confident diagrams and standardized format, and its eventual metric quantification of impact, is not the only or the best vehicle for translating and disseminating today's research: for posing and then answering questions. There's scope for more variety in who does this, and how. Who would have guessed, for example, that what was needed to finally put climate science firmly on the public agenda was the candour and courage of a schoolgirl who is on the autistic spectrum?

The history of science tells us that some of the toughest questions will be addressed not by being answered but by being replaced with better questions. Among those haunting us today that might deserve this fate are: what is life? What is consciousness? What makes individuals who they are? Why does our Universe seem fine-tuned for our existence? How did it all begin? It will take creative and diverse thinking to improve on them – for the view over the horizon might not be the one we anticipated.

#### The author

Philip Ball is a science writer and author; his latest book is How To Grow a Human. e-mail: p.ball@btinternet.com

- 1. Gleiser, M. The Island of Knowledge: The Limits of Science and the Search for Meaning (Basic Books, 2014).
- Mayor, M. & Queloz, D. Nature 378, 355-359 (1995)
- Mermin, N. D. Nature 507, 421-423 (2014)
- Tam, V. et al. Nature Rev. Genet. 20, 467-484 (2019).
- Pennisi, E. Science 362, 1344-1345 (2018).
- Abbott, A. Nature 511, 133-134 (2014) Cobb, M. The Idea of the Brain (Profile, in the press).
- 8. Crawley, A. E. Nature 95, 200-201 (1915)
- Ryle, G. The Concept of Mind (Hutchinson's, 1949)

# Nature's reach: narrow work has broad impact

Alexander J. Gates, Qing Ke, Onur Varol and Albert-László Barabási

A scientific paper today is inspired by more disciplines than ever before, shows a new analysis marking the journal's 150th anniversary.

ow knowledge informs and alters disciplines is itself an enlightening, and vibrant field1. This type of meta research into new findings, insights, conceptual frameworks and techniques is important, among other things, for policymakers who fund research in the hope of tackling society's most pressing challenges, which inevitably span disciplines.

Since its founding in 1869, Nature has offered a venue for publishing major advances from many fields. To mark its anniversary, we track here how papers cite and are cited across disciplines, using data on tens of millions of scientific articles indexed in Clarivate Analytics' Web of Science (WoS), a bibliometric database that encompasses many thousands of research journals starting from 1900. We pay particular attention to articles that appeared in Nature. In our view, this snapshot, for all its idiosyncrasies, reveals how scientific work is ever more becoming a mixture of disciplines.

Several caveats are important. The volatility of our metrics in the early twentieth century can be attributed, at least in part, to the fact that articles then typically had many fewer references and citations. Until the mid-1920s, Nature articles typically listed no references; today, they can have up to 50. Another caveat is that the number of disciplines recognized by WoS grew from 57 in 1900 to 251 in 1993, but this is only one factor contributing to the disciplinary trends we found.

Many scholars have developed methods and metrics to gauge how scientific publishing contributes to knowledge, and to assess influence. For more detailed explanations of our choices, along with essential qualifications, see Supplementary Information (SI).

Across the scientific literature overall, our analysis hints that articles are drawing from and influencing more disciplines than they did 100 years ago, although some disciplines have broader influence than others. As a journal, Nature publishes mostly specialized, or deeply disciplinary, papers; these tend to reference

a narrower range of disciplines than does the average paper. Usually, however, *Nature* papers are cited by a broader range of disciplines than

#### Colossal corpus

We extracted references for papers contained in the WoS publication database from 1900 to 2017, capturing close to 700 million citation relationships. We pinned subsequent analysis to the approximately 19 million articles that had at least one reference and one citation and that were published before 2010 (to give time for citations to accumulate). The resulting corpus integrated the discipline information for 38 million articles.

To identify disciplines, we relied on relatively broad categorizations from WoS. These are necessarily imperfect, but cumulatively reveal patterns of scholarship. Most journals are disciplinary, and so WoS assigns each article to one or more disciplines on the basis of the journal in which it is published. For instance, articles

in the Journal of Bacteriology are categorized as microbiology.

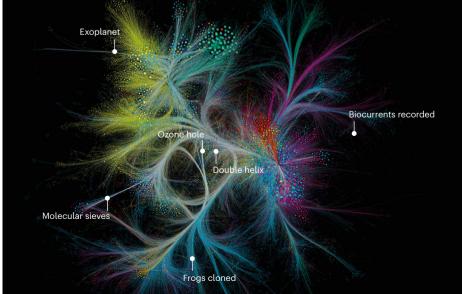
We traced the conceptual journeys to each paper by identifying the inspiration for articles by their references: the works authors credited for their concepts, methods, techniques and insight. Similarly, we identified the impact of each publication by the citations it received in the corpus. Caution is required when using citation-based measures to assess the importance of individual papers or authors; still, the accessibility and quantity of such data provide one view – among many – of how scientific knowledge accumulates1.

We explored how the 88,637 Nature articles in our data set mediate the metabolism of ideas using the broadest WoS disciplinary categories. A Nature article with references mainly from biomedical research will typically collect the largest proportion of its citations from other biomedical-research papers (see 'Knowledge flows'). About half of the papers that cite it will be spread across the other categories. By contrast, a paper with references mainly from engineering and technology is much more likely to be cited by papers in other fields (72%) than by other papers in the same field (28%). Engineering and technology papers also make up a very small proportion of the papers Nature opts to publish; those that are selected might be chosen for their broad appeal. At the other extreme, papers in Earth and space science are much more likely to be cited by papers in their own field (72%) than by other disciplines (28%).

#### CO-CITATION NETWORK

Each Nature paper is a dot. Dots are linked if another paper cites both. Some articles (colourful clusters) are cited by many disciplines, others (monotone areas) are deeply embedded in their own disciplines. (See go.nature.com/n150int for an interactive version, including references to the six highlighted papers.)

- Discipline
- Arts Biology
- Biomedical research Chemistry
- Clinical medicine Earth and space
  - Engineering and technology
- Health
- Humanities
- Mathematics Physics
- Business and management
- Psychology



OURCE DATA: WEB OF SCIENCE. ANALYSIS BY A. J. GATES ET AL.

Another way to reveal intrinsic communities in and across disciplines is through co-citation analysis<sup>2</sup>. In this approach, each paper is represented by a node, shown as a dot. Two papers are linked if another paper cites both of them: the node size reflects the number of co-citations. Our visualization algorithm treats each link as a spring and arranges the nodes to make links as short as possible. This produces clusters of *Nature* papers that vary in their level of interdisciplinary connections (see go.nature. com/n150int).

The overall network structure echoes scientific perceptions of how publications relate to each other. Articles tend to bunch together according to age and topic, because authors usually reference recent articles related to their paper's subject<sup>3</sup>. Over its recent history, more than half of Nature's papers have come from the life sciences. Consequently, clusters of biomedical-research papers appear throughout the network. Since 1930 (when it became reliable to use references to assign papers to disciplines), the proportion of physics papers has shrunk and Earth and space science has grown. Certain papers - such as the discovery of the first exoplanet orbiting a Sun-like star<sup>4</sup> – are deeply embedded in a cluster of papers in the same field. By contrast, the discovery of the ozone hole<sup>5</sup> is in a region where articles of many disciplines - chemistry, social sciences, Earth sciences - are found (see 'Co-citation network'). Our analysis shows that this paper's references are more diverse than 95% of Nature papers, and its citations are more diverse than 99% of Nature papers.

An analysis of the co-citation network from any more-specialized journal would probably look different. Still, distinct episodes from the history of science are apparent in the 3D view of Nature's co-citation network (see go.nature.com/2patums). These include the study of radioactive elements in the 1930s, and how studies of superconducting materials flirted with diverse applications and then were intensely characterized deep within the physical sciences in the late 1980s and 1990s.

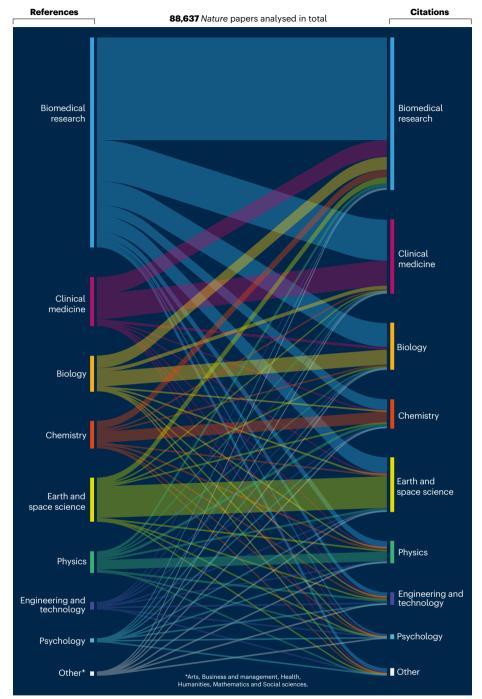
#### Over time

The numbers of papers in every discipline grew exponentially over the past century<sup>1</sup>. Exact rates differ over time, although since about the 1960s, 48% of papers were in the life sciences (with 42% from 'hard' sciences and 10% from behavioural science).

Scholars define and measure influences across disciplines in various ways. Multidisciplinarity usually refers to separate disciplines coming together yet remaining distinct: we define it for journals as the breadth of disciplines that are either inspiring or being impacted by the journal's articles. Interdisciplinarity refers to integration: we define it as the diversity in inspiration in an article's references, and the diversity in how an article's impact diffuses across

#### **KNOWLEDGE FLOWS**

Nature articles are mainly cited by their own disciplines, particularly in some fields, such as Earth and space science. (Each Nature paper was assigned to a discipline using its references, as was every paper in the Web of Science database that cited a Nature paper.)



disciplines. Although it is difficult to assess integration across an article's citations, this measure can capture how the knowledge communicated by the article had diverse impact<sup>6</sup>. This analysis indicates the extent of interactions across disciplines, but does not reveal the specific details of how those disciplines interact.

First, we explored the breadth of disciplines reflected in the references and citations across a journal, capturing the journal's multidisciplinarity (see 'Inspiration and impact'). We labelled each paper in a journal with the primary discipline assigned to its

references (inspiration) or citations (impact), and measured multidisciplinarity on a scale of zero to one. Zero meant that all of an article's references or citations were in the same discipline; one meant that they were balanced evenly across all disciplines, using the normalized entropy measure (see SI). We found that this measure does not depend on the number of articles each journal published (see SI). It probably reflects other qualities of a journal, such as the pool of articles submitted and the editors' selection criteria.

For most journals, the breadth of impact and

The general-science journals *Nature* and *Science* both have a greater breadth of impact (citations) and inspiration (references) than 99.7% of other journals. The multidisciplinarity of *Nature* peaked in the 1960s and has remained relatively high since then, probably reflecting a combination of papers selected by *Nature* that are expected to have broad appeal, and the papers' greater visibility to the scientific community.

Second, we explored the interdisciplinarity of individual articles by measuring the diversity of disciplines in the references and citations<sup>7-10</sup>. Many measures have been proposed to assess interdisciplinarity, and can have inconsistent results (see, for example, refs 11,12). Scholars agree, however, that simply counting the number of disciplines that occur in references and citations is inadequate. For example, a paper that largely references biology and clinical science draws on less diversity than one inspired by biology and physics. We quantify this characteristic on a scale of zero to one using the Rao-Stirling diversity index, which captures the number of disciplines represented, how evenly they are distributed and their degree of difference<sup>13</sup>.

Our analysis shows that the diversity of disciplines in articles' references and citations is increasing. Roughly speaking, a typical article is inspired by and impacts three times more disciplines this decade than it did 50 years ago.

Whereas a typical article published today references articles from the equivalent of 11 disciplines, a *Nature* publication references the equivalent of only 9 (SI, Fig. S5). This is in line with previous analyses suggesting that highly influential work tends to be grounded in deep expertise<sup>14</sup>. By contrast, the disciplinary diversity for the citations of articles in general-science journals has consistently been higher than for articles published elsewhere, suggesting that content in these journals reaches a broader swathe of the scientific community than it drew from. This observation makes sense, considering that these journals aim to reach a broader readership and to publish major advances.

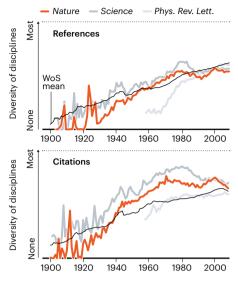
Sometimes, the fields that inspire a paper differ markedly from those on which it has an impact. For example, 'The Digital Code of DNA', a 2003 *Nature* essay by systems biologists Leroy Hood and David Galas<sup>15</sup>, takes most of its inspiration from molecular biology, yet is cited across computer science, clinical medicine and social science. We quantify cross-disciplinarity on a scale from zero to one. In this case, zero implies all disciplines that inspired an article and all those it impacts are identical; a score of one implies these lists differ completely (using the Jensen–Shannon divergence, a measure of the similarity

#### INSPIRATION AND IMPACT

The diversity of disciplines in articles' citations (impact) and references (inspiration) is growing; the likelihood of articles crossing disciplines is not. Articles in *Nature* and *Science* are more broadly cited across disciplines.

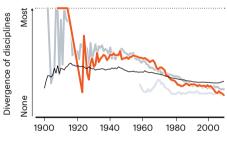
#### Interdisciplinarity

How many, how diverse and how balanced disciplines are across an article's references and citations. This is growing across all of science.



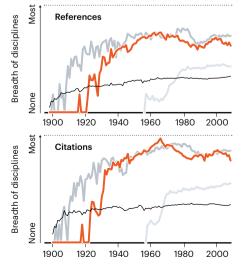
#### **Cross-disciplinarity**

How much the disciplines in articles' references vary from those in their citations. The decline here is probably due to rising interdisciplinarity.



#### Multidisciplinarity

How many disciplines are represented in a journal. Both *Nature* and *Science* consistently show broader impact than the average journal.



Volatile data pre-1930: papers had fewer citations and references, and indexing was less reliable.

between two probability distributions; see SI).

What we see is that in recent decades cross-disciplinarity has declined, with that of the general-science journals falling faster than the scientific literature overall. Perhaps this is because articles that bridge disciplines influence multiple fields, including those from which they arose. As works draw on a broader set of disciplines, there is less scope to influence a set of completely different disciplines.

Assessment of scientific work generally works best when contextualized within its specific discipline. For example, citation counts are more effective when comparing biomedical papers to other biomedical papers rather than to physics papers. But if interactions between disciplines are increasing, then a stringent, coherent assignment makes less sense. We speculate that considering how disciplines intermix within individual articles might allow better comparisons across disciplines or improve assessment of a paper's impact. What's more, strictly structured research departments and funding programmes make less sense if boundaries between disciplines are becoming less distinct. As network scientists, we relish the idea that science is becoming less siloed.

The increase we observe in interdisciplinary thinking is seen across disciplines (see SI) and shows no signs of slowing. With the population of researchers, scientific literature and knowledge ever growing, the scientific endeavour increasingly integrates across boundaries. Research institutions and funding bodies would do well to realize that interdisciplinarity is becoming the norm.

#### The authors

Alexander J. Gates, Qing Ke, Onur Varol and Albert-László Barabási are researchers at the Network Science Institute, Northeastern University, Boston, Massachusetts, USA. e-mail: a.barabasi@northeastern.edu

- Fortunato, S. et al. Science 359, eaao0185 (2018).
- 2. Small, H. J. Am. Soc. Inf. Sci. 24, 265–269 (1973).
- Mukherjee, S., Romero, D. M., Jones, B. & Uzzi, B. Sci. Adv. 3, e1601315 (2017).
- 4. Mayor, M. & Queloz, D. Nature 378, 355-359 (1995).
- Farman, J. C., Gardiner, B. G. & Shanklin, J. D. Nature 315, 207–210 (1985).
- Leydesdorff, L., Wagner, C. S. & Bornmann, L. Scientometrics 114, 567–592 (2018).
- Choi, B. C. K. & Pak, A. W. P. Clin. Invest. Med. 29, 351–364 (2006).
- 8. Porter, A. L. & Rafols, I. Scientometrics 81, 719 (2009).
- 9. Wagner, C. S. et al. J. Informetr. 5, 14–26 (2011).
- Leydesdorff, L. & Rafols, I. J. Informetr. 5, 87–100 (2011).
   Wang, Q. & Schneider, J. W. Preprint at https://arxiv.org/abs/1810.00577 (2018).
- Research Councils UK & Digital Science. Interdisciplinary Research: Methodologies for Identification and Assessment (RCUK/Digital Science, 2016).
- 13. Stirling, A. J. R. Soc. Interface 4, 707-719 (2007)
- 14. Uzzi, B., Mukherjee, S., Stringer, M. & Jones, B. *Science* **342**. 468–472 (2013).
- 15. Hood, L. & Galas D. Nature 421, 444-448 (2003)

Supplementary information accompanies this article: see go.nature.com/2wtoux3.

# 10 extraordinary papers

#### Genetics

### The structure of DNA

#### **Georgina Ferry**

In the early 1950s, the identity of genetic material was still a matter of debate. The discovery of the helical structure of double-stranded DNA settled the matter – and changed biology forever.

On 25 April 1953, James Watson and Francis Crick announced1 in Nature that they "wish to suggest" a structure for DNA. In an article of just over a page, with one diagram (Fig. 1), they transformed the future of biology and gave the world an icon - the double helix. Recognizing at once that their structure suggested a "possible copying mechanism for the genetic material", they kick-started a process that, over the following decade, would lead to the cracking of the genetic code and, 50 years later, to the complete sequence of the human genome.

Until that time, biologists had still to be convinced that the genetic material was indeed DNA; proteins seemed a better bet. Yet the evidence for DNA was already available. In 1944, the Canadian-US medical researcher Oswald Avery and his colleagues had shown<sup>2</sup> that the transfer of DNA from a virulent to a non-virulent strain of bacterium conferred virulence on the latter. And in 1952, the biologists Alfred Hershev and Martha Chase had published evidence<sup>3</sup> that phage viruses infect bacteria by injecting viral DNA.

Watson, a 23-year-old US geneticist, arrived at the Cavendish Laboratory at the University of Cambridge, UK, in autumn 1951. He was convinced that the nature of the gene was the key problem in biology, and that the key to the gene was DNA. The Cavendish was a physics lab, but also housed the Medical Research Council's Unit for Research on the Molecular Structure of Biological Systems, headed by chemist Max Perutz. Perutz's group was using X-ray crystallography to unravel the structures of the proteins haemoglobin and myoglobin. His team included a 35-year-old graduate student who had given up physics and retrained in biology, and who was much happier working out the theoretical implications of other people's results than doing experiments of his own: Francis Crick. In Crick, Watson found a ready ally in his DNA obsession.

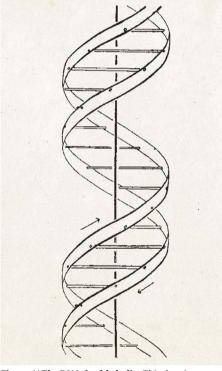


Figure 1 | The DNA double helix. This drawing appeared in Watson and Crick's report1 of the structure of DNA, and was produced by Crick's wife, Odile.

However, DNA was the project of Maurice Wilkins at King's College London. Crick was a friend of Wilkins's, and it wasn't the done thing for labs to compete over the same molecule. Moreover, the experienced X-ray crystallographer Rosalind Franklin had just taken over experimental work on DNA at King's. Owing to a misunderstanding about their relative roles, Franklin's relationship with Wilkins was frosty.

None of this stopped Watson and Crick from speculating about how the components of the DNA molecule – the four nucleotide bases adenine, guanine, thymine and cytosine, connected to a backbone of sugars and phosphates - might assemble into fibres. They thought that a helix was a likely option: the US chemist Linus Pauling and his co-workers had just demonstrated4 that peptide chains formed α-helices. Crick himself had co-authored a paper on the theory of diffraction of X-rays by helices5. In late 1951, he and Watson combined that theory with what they knew about the chemistry of DNA, and what they remembered of talks given by Wilkins and Franklin, to build a model of the DNA structure.

They got it badly wrong: Wilkins and Franklin quickly demolished it. The head of the Cavendish, Lawrence Bragg, was furious, and banned Watson and Crick from doing any further work on DNA. But then, in February 1952, the Cavendish team received a manuscript from Pauling that contained a DNA model. It was wrong, but Watson and Crick were alarmed that Pauling was potentially near a solution.

This time, Bragg agreed that they might try to get there first. Franklin was soon to move to Birkbeck College, London, and was leaving the DNA work to Wilkins. She and her graduate student, Raymond Gosling, had given Wilkins a photograph of the X-ray-diffraction pattern produced by the B form of DNA. Watson went to see Wilkins, who showed him the photograph, without Franklin and Gosling's knowledge.

The now famous 'Photograph 51', together with other unpublished data of Franklin's that Perutz had shown Watson and Crick, told the pair that DNA did indeed form a helix, and that the structure consisted of two chains running in opposite directions. Watson was stumped, however, over how the bases could pair up between the two. He made cardboard cutouts of the bases, trying to fit them together, but nothing seemed to work.

His colleague Jerry Donohue then pointed out that he was using the molecular structures of the enol isomers of the bases, which cannot form the hydrogen bonds necessary for base-pairing. Once Watson had made cutouts of the alternative keto isomers, he had the blinding revelation that when guanine bonded to cytosine, it made an identical shape to that of adenine bonded to thymine, and that the shapes fitted perfectly into the helical frame provided by the backbones of each DNA chain. This explained biochemist Erwin Chargaff's discovery that the DNA of any species has the same amount of guanine as of cytosine, and of adenine as of thymine<sup>6</sup>. It also showed that each DNA chain in a helix provides a perfect template for the other, reading the base sequence in opposite directions.

#### 10 extraordinary papers

Within days, Watson and Crick had built a new model of DNA from metal parts. Wilkins immediately accepted that it was correct. It was agreed between the two groups that they would publish three papers simultaneously in *Nature*, with the King's researchers commenting on the fit of Watson and Crick's structure to the experimental data, and Franklin and Gosling publishing Photograph 51 for the first time<sup>7,8</sup>.

The Cambridge pair acknowledged in their paper that they knew of "the general nature of the unpublished experimental results and ideas" of the King's workers, but it wasn't until *The Double Helix*, Watson's explosive account of the discovery, was published in 1968 that it became clear how they obtained access to those results. Franklin had died of cancer a decade previously; her death prevented her from sharing the Nobel prize awarded to Watson, Crick and Wilkins in 1962.

The immediate reception of the double-helix model was surprisingly muted<sup>9</sup>, perhaps because there was no obvious mechanism to explain its role in protein synthesis. In a landmark talk in 1957, Crick proposed that the base sequence encoded the sequence of amino acids in a protein, and that protein production involved RNA both as a template and as an 'adaptor' that would enable amino acids to be attached to one another in the right order. He also supported the suggestion originally made informally by the physicist George Gamow to the members of the 'RNA Tie Club' convened by Gamow and Watson, but also independently proposed by biologist Sydney Brenner<sup>10</sup> – that triplets of bases (which Brenner called codons) encode the 20 amino acids commonly found in proteins. Finally, Crick expounded what he called the 'central dogma' of biology: that information can flow from nucleic acids to proteins, but not the other way round11.

These predictions were confirmed by experiment in the next few years. In 1958, the biochemists Matthew Meselson and Franklin Stahl showed that one DNA strand acts as a template for the formation of a new strand<sup>12</sup>. The same year, Arthur Kornberg and his colleagues published their discovery of the enzyme DNA polymerase<sup>13</sup>, which adds bases to newly forming strands. Messenger RNA, transfer RNA and ribosomal RNA were all quickly identified.

In 1961, Marshall Nirenberg and Heinrich Matthaei were the first to crack part of the genetic code, demonstrating that bacterial extracts synthesize only the amino acid phenylalanine from RNA that contains just one type of RNA base<sup>14</sup> (uracil; U). The same year, Crick, his indispensable female technician Leslie Barnett and their co-workers reported mutation studies that confirmed the existence of the triplet-based code<sup>15</sup>, and which therefore suggested that the codon for phenylalanine was UUU. The race to

identify the full set of codons was completed by 1966, with Har Gobind Khorana contributing the sequences of bases in several codons from his experiments with synthetic polynucleotides (see go.nature.com/2hebk3k).

With Fred Sanger and colleagues' publication of an efficient method for sequencing DNA in 1977, the way was open for the complete reading of the genetic information in any species. The task was completed for the human genome by 2003, another milestone in the history of DNA.

Watson devoted most of the rest of his career to education and scientific administration as head of the Cold Spring Harbor Laboratory in Long Island, New York, and serving (briefly) as the first head of the US National Center for Human Genome Research, now the National Human Genome Research Institute. Always outspoken, he was eventually removed from his emeritus position at Cold Spring Harbor when he repeatedly aired controversial opinions about genetics, race and intelligence.

Crick continued to tackle hard problems in science, moving in 1977 from Cambridge to the Salk Institute in La Jolla, California, where he spent the rest of his life working on the neural basis of consciousness<sup>17</sup> and, specifically, of visual perception. He died in 2004, aged 88.

The double helix put genetics on a physical footing that would shed light on almost every aspect of modern biology and medicine. Examples include the migration of human populations throughout history; ecology and biodiversity; cancer-causing mutations in tumours and their drug treatment; surveillance of microbial drug resistance in hospitals and the global population; and the diagnosis and treatment of rare congenital diseases. DNA analysis has long been established

in forensics, and research into more-futuristic applications, such as DNA-based computing, is well advanced.

Paradoxically, Watson and Crick's iconic structure has also made it possible to recognize the shortcomings of the central dogma, with the discovery of small RNAs that can regulate gene expression, and of environmental factors that induce heritable epigenetic change. No doubt, the concept of the double helix will continue to underpin discoveries in biology for decades to come.

**Georgina Ferry** is a science writer based in Oxford, UK. A revised edition of her biography *Dorothy Crowfoot Hodgkin* has just been published by Bloomsbury Reader.

- Watson, J. D. & Crick, F. H. C. Nature 171, 737–738 (1953).
- Avery, O. T., MacLeod, C. M. & McCarty, M. J. Exp. Med. 79, 137–158 (1944).
- Hershey, A. D. & Chase, M. J. Gen. Physiol. 36, 39–56 (1952).
- Pauling, L., Corey, R. B. & Branson, H. R. Proc. Natl Acad. Sci. USA 37, 205–211 (1951).
- Cochran, W., Crick, F. H. & Vand, V. Acta Crystallogr. 5, 581–586 (1952).
- Vischer, E. & Chargaff, E. J. Biol. Chem. 176, 703-714 (1948).
- Wilkins, M. H. F., Stokes, A. R. & Wilson, H. R. Nature 171, 738–740 (1953).
- 8. Franklin, R. E. & Gosling, R. G. Nature 171, 740–741 (1953).
- 9. Olby, R. Nature 421, 402-405 (2003).
- Brenner, S. Proc. Natl Acad. Sci. USA 43, 687–694 (1957).
   Crick, F. H. C. Symp. Soc. Exp. Biol. 12, 138–163 (1958).
- Meselson, M. & Stahl, F. W. Proc. Natl Acad. Sci. USA 44, 671–682 (1958).
- Lehman, I. R., Bessman, M. J., Simms, E. S. & Kornberg, A. J. Biol. Chem. 233, 163–170 (1958).
- Nirenberg, M. W. & Matthaei, J. H. Proc. Natl Acad. Sci. USA 47, 1588–1602 (1961).
- Crick, F. H. C., Barnett, L., Brenner, S. & Watts-Tobin, R. J. Nature 192, 1227–1232 (1961).
- Sanger, F., Nicklen, S. & Coulson, A. R. Proc. Natl Acad. Sci. USA 74, 5463–5467 (1977).
- 17. Crick, F. H. C. The Astonishing Hypothesis: The Scientific Search for the Soul (Simon & Schuster, 1994).

#### **High-energy physics**

# Detection of a strange particle

#### Taku Yamanaka

In 1947, scientists found a previously unseen particle, which is now called a neutral kaon. This work led to the discovery of elementary particles known as quarks, and ultimately to the establishment of the standard model of particle physics.

In the late 1940s, the physicists George Rochester and Clifford Butler<sup>1</sup> observed something unusual in their charged-particle detector. They were studying the interactions between high-energy cosmic rays and a lead plate in the detector when they spotted V-shaped particle tracks (Fig. 1a). The small

gap between the lead plate and the vertex of the tracks indicated that an invisible neutral particle had been produced in the plate, had travelled for a short distance and had then decayed into two visible charged particles. The mass of the neutral particle was about 1,000 times that of an electron, implying

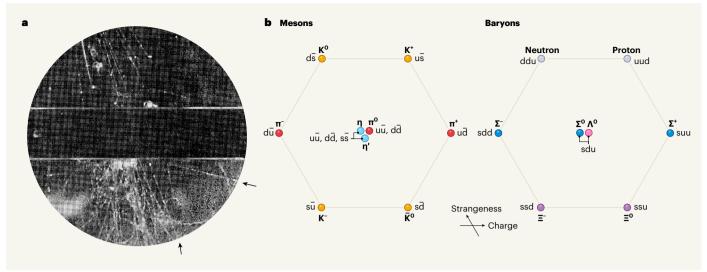


Figure 1 | Particle detection that led to a better understanding of fundamental physics. a, In 1947, Rochester and Butler<sup>1</sup> analysed the particles produced when high-energy cosmic rays hit a lead plate (the broad central stripe) in a charged-particle detector. In certain photographs, they spotted evidence of a previously undetected, invisible neutral particle decaying into two visible charged particles, which were identified by tracks (labelled with arrows). b, The discovery

of many more particles following Rochester and Butler's work led to a model<sup>12,13</sup> in which all of the known mesons and baryons (two classes of particle) consist of elementary particles called up (u), down (d) and strange (s) quarks, along with their antiparticles (denoted by overbars). The  $\eta$ ,  $\eta'$  and  $\pi^0$  mesons comprise mixtures of quark pairs. The mesons and baryons are arranged by their strangeness (a quantity that is related to the presence of strange quarks) and electric charge.

that it must be a previously unreported type of particle. This discovery paved the way for many puzzles and surprises in particle physics in the decades that followed.

At the time of Rochester and Butler's work, protons, neutrons, electrons and particles called pions (short for  $\pi$  mesons) had been identified, and were known to be sufficient to form atoms. Pions were proposed<sup>2</sup> in 1935 to explain how protons and neutrons are held together in small atomic nuclei by the strong nuclear force, and were found experimentallv<sup>3,4</sup> in 1947.

While searching for a pion in cosmic rays, scientists discovered a different particle<sup>5</sup>. which is now called a muon. A heavy charged particle was then found in 1944, followed by Rochester and Butler's unstable neutral particle. But the discovery of unexpected particles did not stop there. Then came the τ meson, which decays into three pions; the  $\theta$  meson, decaying into two pions; the k meson, decaying into a muon and an invisible particle; the  $\Lambda^0$  particle, decaying into a proton and a pion; and the list goes on.

In the early 1950s, researchers began producing these rare particles in large quantities by firing protons at targets in particle accelerators. The  $\tau$ ,  $\theta$  and  $\kappa$  mesons and  $\Lambda^0$  particle were peculiar, because, although they were generated by the strong force, their decay times were much longer than those expected for this force. To explain these observations, physicists proposed a quantity, known as strangeness (S), that is conserved by the strong force<sup>7,8</sup>.

Protons and neutrons have S = 0, and through the strong force, can produce a pair of strange particles that have S = -1 and S = +1, so that total strangeness is conserved. However, a strange particle that has S = -1, for example, cannot decay into particles that have S = 0 through the strong force, because strangeness would not be conserved. Instead, this decay must occur much more slowly through the weak nuclear force, which allows total strangeness to change.

As the accuracy of accelerator-based measurements increased, it became clear that the  $\tau$  and  $\theta$  mesons had extremely similar masses and lifetimes. Scientists concluded that these mesons must be the same particle, which is able to decay into two or three pions. The mess of strange mesons was finally cleaned up into four particles dubbed kaons (short for K mesons): K<sup>+</sup> and K<sup>0</sup> and their antiparticles  $K^-$  and  $\bar{K}^0$ .

However, accepting that the  $\tau$  and  $\theta$  mesons were the same particle raised another problem. A state of two pions has even parity, which means that its wavefunction does not change sign under a parity transformation (in which spatial coordinates are flipped). By contrast, a state of three pions has odd parity. If the same particle could decay into two or three pions, did that mean that, contrary to all conventional wisdom, parity is not conserved by the weak force? This question, known as the  $\tau$ - $\theta$  puzzle, led to the discovery, in 1957, of such parity-symmetry breaking in cobalt-60 decays9 and in pion decays10.

A consequence of parity-symmetry breaking by the weak force is that elementary particles called neutrinos can be only left-handed, which means that their motion and intrinsic angular momentum are in opposite directions. Under a parity transformation, a left-handed neutrino becomes a right-handed neutrino, which does not exist. However, if one then applies a charge-conjugation transformation (in which particles are replaced by their antiparticles), the right-handed neutrino becomes a right-handed antineutrino, which does exist. The weak force therefore seemed to conserve CP symmetry (symmetry under a combined charge-conjugation and parity transformation), until such symmetry was found to be broken in neutral-kaon decays.

A neutral kaon is a mixture of  $K^0$  and  $\bar{K}^0$ states, and can exist as the CP-even state K<sub>even</sub> or the CP-odd state K<sub>odd</sub>. The lifetime of  $K_{odd}$  is much longer than that of  $K_{even}$ , so these particles were named K<sub>1</sub> (for 'K-long') and K<sub>5</sub> (for 'K-short'), respectively. A useful consequence of such lifetimes is that, if neutral kaons are produced by firing protons at a target, the CP-even K<sub>s</sub> component quickly decays, leaving only the CP-odd K<sub>1</sub> component. In 1964, such K<sub>1</sub> particles were observed<sup>11</sup> to decay into the CP-even state of two oppositely charged pions  $(\pi^+\pi^-)$ . Therefore, despite expectations, CP symmetry was shown to be broken.

In that same year, physicists proposed a  $model^{12,13}\,to\,explain\,all\,of\,the\,known\,mesons$ and baryons — a family that includes protons, neutrons and the  $\Lambda^0$  particle. In the model, these mesons and baryons consist of elementary particles known as quarks, which come in three types: up, down and strange (Fig. 1b).

In 1973, a theoretical model14 showed that the breaking of CP symmetry could be explained by introducing three more quarks: charm, top and bottom. In this framework, K<sub>1</sub> can have a small component of  $K_{\text{even}}$  that can decay into the CP-even  $\pi^+\pi^-$  state. But unlike other theoretical models, this framework also allows Kodd to decay into the CP-even state (direct CP violation).

#### 10 extraordinary papers

Many generations of experiments then were carried out to see whether direct CP violation exists. The measurement required extremely high precision, and after many improvements over 25 years, direct CP violation was finally confirmed <sup>15,16</sup>. Together with the observation of CP-symmetry breaking in B mesons (mesons that contain a bottom quark) <sup>17,18</sup>, the theoretical model was confirmed, and helped to establish the standard model of particle physics, which is the current explanation of the Universe's particles and forces.

However, the standard model is not complete. For instance, it cannot explain why the Universe contains so little antimatter, nor what the mysterious substance called dark matter is. Researchers are therefore trying to search for a hint of particle physics beyond that of the standard model. For example, experiments in Japan<sup>19</sup> and Europe<sup>20</sup> are using extremely rare kaon decays to search for such a hint.

In retrospect, Rochester and Butler's V-shaped particle tracks are thought to have been caused by a  $K_s$ , produced in the lead plate, decaying into the  $\pi^+\pi^-$  state. Since their work, kaons have been used to discover strangeness and the breaking of parity and CP symmetries, to build the quark model and the standard model, and now to search for previously unseen particle physics. Could Rochester and Butler have ever imagined that they had opened such a treasure chest?

**Taku Yamanaka** is in the Department of Physics, Osaka University, Toyonaka, Osaka 560-0043, Japan.

e-mail: taku@champ.hep.sci.osaka-u.ac.jp

- Rochester, G. D. & Butler, C. C. Nature 160, 855–857 (1947).
- Yukawa, H. Proc. Phys.-Math. Soc. Jpn 17, 48–57 (1935).
- 3. Perkins, D. H. Nature 159, 126-127 (1947).
- Lattes, C. M. G., Muirhead, H., Occhialini, G. P. S. & Powell, C. F. Nature 159, 694–697 (1947).
- Neddermeyer, S. H. & Anderson, C. D. Phys. Rev. 51, 884–886 (1937).
- Leprince-Ringuet, L. & L'Héritier, M. C.R. Acad. Sci. 219, 618–620 (1944).
- 7. Gell-Mann, M. Phys. Rev. 92, 833-834 (1953).
- Nakano, T. & Nishijima, K. Prog. Theor. Phys. 10, 581–582 (1953).
- Wu, C. S., Ambler, E., Hayward, R. W., Hoppes, D. D. & Hudson, R. P. Phys. Rev. 105, 1413–1415 (1957).
- Garwin, R. L., Lederman, L. M. & Weinrich, M. Phys. Rev. 105, 1415–1417 (1957).
- Christenson, J. H., Cronin, J. W., Fitch, V. L. & Turlay, R. Phys. Rev. Lett. 13, 138–140 (1964).
- 12. Gell-Mann, M. Phys. Lett. **8**, 214–215 (1964).
- 13. Zweig, G. CERN Rep. No. CERN-TH-401 (1964).
- Kobayashi, M. & Maskawa, T. Prog. Theor. Phys. 49, 652–657 (1973).
- Alavi-Harati, A. et al. (KTeV Collaboration) Phys. Rev. Lett. 83, 22–27 (1999).
- The NA48 Collaboration. Eur. Phys. J. C 22, 231–254 (2001)
- 17. Aubert, B. et al. (BABAR Collaboration) Phys. Rev. Lett. 87, 091801 (2001)
- Abe, K. et al. (Belle Collaboration) Phys. Rev. Lett. 87, 091802 (2001).
- Ahn, J. K. et al. (KOTO Collaboration) Phys. Rev. Lett. 122, 021802 (2019).
- 20. The NA62 Collaboration. *Phys. Lett. B* **791**, 156–166 (2019).

#### **Neuroscience**

# Neuronal signals thoroughly recorded

#### Alexander D. Reyes

Originally developed to record currents of ions flowing through channel proteins in the membranes of cells, the patch-clamp technique has become a true stalwart of the neuroscience toolbox.

Information in the brain is thought to be encoded as complex patterns of electrical impulses generated by thousands of neuronal cells. Each impulse, known as an action potential, is mediated by currents of charged ions flowing through a neuron's membrane. But how the ions pass through the insulated membrane of the neuron remained a puzzle for many years. In 1976, Erwin Neher and Bert Sakmann developed the patch-clamp technique, which showed definitively that currents result from the opening of many channel proteins in the membrane<sup>1</sup>. Although the technique was originally designed to record tiny currents, it has since become one of the most important tools in neuroscience for studying electrical signals - from those at the molecular scale to the level of networks of neurons.

By the 1970s, current flowing through the cell was generally accepted to result from the opening of many channels in the membrane, although the underlying mechanism was unknown. At that time, current was commonly recorded by impaling tissue with a sharp electrode — a pipette with a very fine point. Unfortunately, however, the signal recorded in this way was excessively noisy, and so only the large, 'macroscopic' current — the collective current mediated by many different types of channel — that flows through the tissue could be resolved.

In 1972, Bernard Katz and Ricardo Miledi<sup>2</sup>, pioneers of the biology of the synaptic connections between cells, managed to infer from the macroscopic current certain properties of the membrane channels, but only after a heroic effort to exclude all possible confounding factors. The problem was that the macroscopic current could be influenced by factors not directly related to channel activity, such as cell geometry and modulatory processes that regulate cell excitability. Also troublesome was that interpretations of macroscopic-current features were based on unverified assumptions about the statistics of individual channel activity<sup>2,3</sup>. Despite Katz and Miledi's careful analyses, there was a lingering doubt about whether their conclusions were correct. The crucial data

were obtained by Neher and Sakmann using patch clamp.

The patch-clamp technique is conceptually rather simple. Instead of impaling the cells, a pipette with a relatively large diameter is pressed against the cell membrane. Under the right conditions, the pipette tip 'bonds' with the membrane, forming a tight seal. This substantially reduces the noise compared with that encountered using sharp electrodes, because the small patch of membrane encompassed by the pipette tip is electrically isolated from the rest of the cell's membrane and from the environment surrounding the cell (Fig. 1).

The tiny currents passing through the few channels in the patch were thus observed for the first time. The recording confirmed key channel properties: when channels open, there is a step-like jump in the current trace and, when they close, a step-like drop back to baseline. It was now possible to determine details such as the statistics of the opening and closing of channels, the amplitude of the currents they mediate and the optimal stimuli that trigger their opening. For this work, Neher and Sakmann were awarded the 1991 Nobel Prize in Physiology or Medicine.

Improvements in patch clamp made it feasible to study channels in various preparations<sup>4</sup> to finally address long-standing questions. There was particular interest in verifying a model for action-potential generation<sup>5</sup> proposed by Nobel laureates Alan Hodgkin and Andrew Huxley in the 1950s. Specific predictions of the model could now be tested directly by examining the current through individual channels and by observing the changes in current that occur when the molecular structure of the channel is modified<sup>6</sup>. Ultimately, the model was shown to be mostly correct and remains the gold standard for computational neuroscientists today.

One of the several variants of patch clamp<sup>4</sup> – the whole-cell configuration – found an audience with neuroscientists studying electrical phenomena in neurons beyond the channel level. To achieve whole-cell recording, the

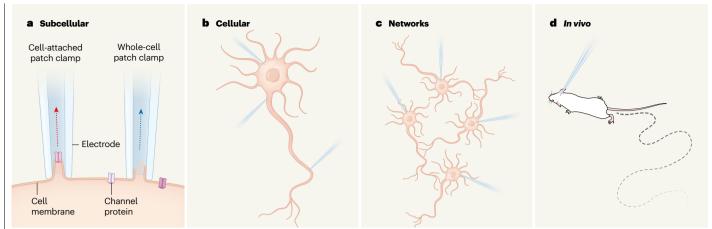


Figure 1 | The patch-clamp technique used at different scales. a, Neher and Sakmann<sup>1</sup> developed the cell-attached patch-clamp technique. An electrode (a fine pipette) is pressed against a 'patch' of the cell membrane so that ion currents (red dotted arrow) passing through channel proteins in the patch under the electrode can be recorded. In the whole-cell configuration, the patch is ruptured so that the whole-cell macroscopic current (blue dotted

arrow), which represents the summed currents from the entire cell, can be recorded. b, Simultaneous whole-cell recordings from different parts of a neuron can determine, for example, the direction of travelling signals.  $\boldsymbol{c}\text{, Whole-cell recordings can be made from a small network of connected}$ neurons. d. Whole-cell recording can even be made in the brains of animals performing a task or walking around freely.

patch of membrane under the electrode is ruptured, enabling electrical access to the cell. Compared with the use of sharp electrodes, whole-cell patch clamp allows much more accurate recordings and, crucially, is less damaging to the cell. This allowed systematic investigation of synergistic processes at the cellular level, such as the regulation of macroscopic currents by modulatory molecules, and interactions between the different types of channel in the neuron

The relatively large opening created in the cell in the whole-cell configuration also provided access to the cell by chemicals, enabling dyes to be delivered for visualizing intricate cell structures, and RNA to be extracted for gene-expression analysis<sup>7</sup>. Neher's group examined the sequence of events that underlie the transfer of information between cells by introducing chemicals into the cell and simultaneously tracking the resulting changes in the electrical properties of the cell's membrane<sup>8</sup>.

Whole-cell patch clamp proved ideal for studying the collective properties of neurons and neuronal networks in brain slices maintained in vitro. A challenge in working with more-complex systems such as neuronal networks is that the number of possible confounding factors increases. Sakmann's solution in the 1990s was to carry out simultaneous whole-cell recording using two or three electrodes, which to some seemed excessive because comparable data could be obtained by sequential recordings using fewer electrodes. However, the rationale was that taking time to design the near-perfect experiment mitigated later difficulties in data interpretation analogous to those faced by Katz and Miledi.

Hence, simultaneous recordings from different parts of the neuron definitively confirmed that action potentials are initiated at one part of the main long neuronal protrusion (the axon) and propagate back to the dendrites (clustered protrusions that receive inputs from other neurons)9. The mechanisms that underlie signalling between neurons were directly investigated by placing electrodes on either side of a synaptic connection<sup>10</sup>. Moreover, triple recordings from neurons of different classes uncovered certain basic principles of network organization<sup>11</sup>.

The patch-clamp technique is also used to examine cell activities under more natural conditions. To study how sensory stimuli and movements are represented in the brain, experiments must be carried out in living animals. The challenge with this approach, however, is that the slightest movement can dislodge an electrode from the neuron. Wholecell patch-clamping turns out to be remarkably stable because of the tight seal between

#### "Patch-clamp recording is arguably still the most effective way of studying electrical signals in the brain."

the electrode and the membrane. Thus, this technique has permitted recording from dendrites<sup>12</sup> and pairs of neurons<sup>13</sup> in anaesthetized rodents, and even from animals that are able to walk and run<sup>14</sup>.

Patch-clamp recording is arguably still the most direct and effective way of studying electrical signals in the brain. The data obtained with this technique essentially represent the ground truth for investigators in many branches of neuroscience, from theorists<sup>15</sup> to translational researchers developing drugs for the treatment of certain brain conditions, including epilepsy<sup>16</sup> and autism spectrum disorder<sup>17</sup>.

Moreover, patch clamp complements modern 'optogenetic' techniques, which enable control and visualization of the activities of large populations of neurons using light18. Emerging technologies, such as prostheses for vision19, will probably rely heavily on patch-clamp recording to establish the optimal conditions for converting external stimuli into electrical signals. Patch-clamping will clearly remain a vital tool for the neuroscientist in the foreseeable future.

Alexander D. Reyes is in the Center for Neural Science, New York University, New York, New York 10003, USA. e-mail: ar65@nyu.edu

- Neher, E. & Sakmann, B. Nature 260, 799-802 (1976)
- Katz, B. & Miledi, R. J. Physiol. (Lond.) 224, 665-699 (1972).
- Anderson, C. R. & Stevens, C. F. J. Physiol. (Lond.) 235, 655-691 (1973).
- Hamill, O. P., Marty, A., Neher, E., Sakmann, B. & Sigworth, F. J. Pflügers Arch. Ges. Physiol. 391, 85-100 (1981). Hodgkin, A. L. & Huxley, A. F. J. Physiol. (Lond.) 117,
- 500-544 (1952).
- Ahern, C. A., Payandeh, J., Bosmans, F. & Chanda, B. J. Gen. Physiol. 147, 1-24 (2016).
- Wang, Y., Gupta, A., Toledo-Rodriguez, M., Wu, C. Z. & Markram, H. Cereb. Cortex 12, 395-410 (2002).
- Neher, E. & Marty, A. Proc. Natl Acad. Sci. USA 79, 6712-6716 (1982).
- 9. Stuart, G. J. & Sakmann, B. Nature 367, 69-72 (1994).
- 10. Borst, J. G., Helmchen, F. & Sakmann, B. J. Physiol, (Lond.) 489, 825-840 (1995).
- Reves. A. et al. Nature Neurosci. 1, 279–285 (1998).
- 12. Smith, S. L., Smith, I. T., Branco, T. & Häusser, M. Nature 503, 115-120 (2013).
- 13. Jouhanneau, J.-S. & Poulet, J. F. A. Front. Synaptic Neurosci, 11, 15 (2019).
- 14. Lee, A. K., Manns, I. D., Sakmann, B. & Brecht, M. Neuron 51, 399-407 (2006).
- 15. Barral, J. & Reyes, A. D. Nature Neurosci. 19, 1690-1696 (2016).
- 16. Catterall, W. A. Annu. Rev. Pharmacol. Toxicol. 54, 317–338
- 17. Daghsni, M. et al. Brain Behav. 8, e00978 (2018).
- 18. Kim, C. K., Adhikari, A. & Deisseroth, K. Nature Rev. Neurosci. 18, 222-235 (2017).
- 19. Berry, M. H. et al. Nature Commun. 10, 1221 (2019).

Materials science

# Birth of a class of nanomaterial

#### **Ryong Ryoo**

Nearly 30 years ago, a simple chemical principle was reported that enabled the synthesis of a plethora of porous materials — some of which might enable applications ranging from biomedicine to petrochemical processing.

In 1992, Kresge et al.¹ reported a breakthrough in materials science. They described multimolecular templates that guide the assembly of ordered mesoporous molecular sieves — materials that contain uniform, regularly arranged pores with mesoscopic diameters (between 2 and 50 nanometres). Their findings triggered an explosion of research into mesoporous materials, which have since been intensively studied for applications as diverse as catalysis, molecular adsorption, drug delivery and molecular separations using membranes.

When Kresge and colleagues published their work, materials known as zeolites — crystalline aluminosilicate compounds with uniform pores usually less than 2 nm in diameter — had long been used as catalysts in petroleum refining and for molecular separations. However, large molecules, such as those found in the heavy fractions of crude oil, were unable to diffuse through the small pores of these molecular sieves, and so could not be processed efficiently. There had been many attempts to obtain zeolite-like materials with enlarged pores and ordered structures, but the large-pore materials commonly available at that time all had a broad

distribution of pore diameters, making them unsuitable for many applications.

In 1990, it was reported<sup>2</sup> that the spaces between the layers of a silicate material called kanemite could be expanded by adding organic molecules containing long hydrocarbon chains to a suspension of kanemite powder in water. This process could generate pores of up to 4 nm in diameter, but worked only with kanemite.

Enter Kresge and colleagues, whose method for making mesoporous materials began with the formation of layers of silica, a few nanometres thick, in between the surfaces of cylindrical supramolecular assemblies called micelles (Fig. 1). The micelles consisted of numerous detergent-like molecules, known as surfactants, that were packed together to form a liquid-crystal structure reminiscent of a honeycomb. Once silica layers had formed between the micelles, the researchers heated the resulting material in air to remove the surfactant, thereby producing a silica product that retained a honevcomb-like array of nanometre-scale pores. The researchers named their material Mobil Composition of Matter No. 41 (MCM-41), after the oil company that they worked for.

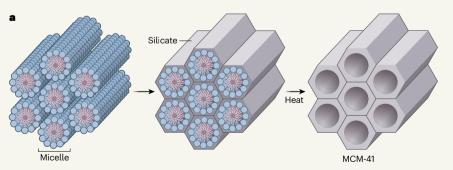
The most impressive feature of Kresge and

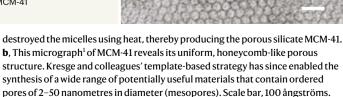
colleagues' strategy was that the diameter, shape and connectivity of the pores could, in principle, be controlled by manipulating the structure and size of the surfactant molecules. The authors demonstrated only a few examples of this: they showed that the pore diameter could be controlled within a narrow range of about 3–10 nm. Nevertheless, their approach was later shown to be applicable to the full range of mesopore sizes<sup>3</sup>.

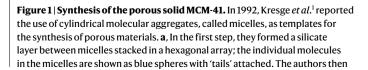
Researchers in the field initially regarded Kresge and colleagues' work as simply extending the pore sizes of the existing family of molecular sieves. However, it soon became apparent that the surfactant-based strategy could be used to synthesize many types of ordered mesoporous material, including ones made from metal oxides<sup>4</sup>, organic polymers<sup>5</sup> and even transition metals<sup>6</sup>. Having the ability to make a variety of mesoporous materials that contain highly ordered arrangements of pores opened up many avenues of research for nanoscience.

A key development in 1998 was the use of polymeric surfactants<sup>7</sup>, which increased the size of mesopores that could be made to 30 nm. Polymeric surfactants used in the synthesis of such large-pored materials, as well as other organic surfactants (including the one used to make MCM-41) are now classified as soft templates, which reflects the somewhat deformable nature of the micelles that act as the mould. Anadvantage of using soft templates is that mesoporous materials can be made in solution at relatively low temperatures. Moreover, the porous structure of the resulting material can easily be controlled by making simple modifications to the template molecules.

Another breakthrough, reported in 1999, was the discovery of hard templating (also known as nanocasting)<sup>8,9</sup>. In this process, mesoporous materials are fabricated from precursor molecules using another solid mesoporous material as a mould, in a manner analogous to the casting of concrete pipes or bricks. Nanocasting has two somewhat cumbersome requirements:







the precursors must infiltrate the pores of the mould uniformly, without accumulating on the external surface; and the precursors must convert completely into the desired product. The method does, however, work particularly well when high temperatures (of the order of 500 °C or more) are needed to synthesize a mesoporous material. This contrasts with the use of surfactant-based soft templates, which typically decompose at temperatures above 200 °C.

Nanocasting was first used to make ordered mesoporous carbon<sup>8</sup>, but has since been developed as a general approach for synthesizing nanowires and nanoporous materials of various compositions, including metal oxides, organic polymers and metals<sup>10</sup>. Mesoporous carbons have garnered much interest because of their high electrical conductivity11, and because they can accommodate a large volume of guest atoms, molecules or particles inside the mesopores. For this reason, mesoporous carbons are considered to be particularly attractive candidates for electrode materials in chemical sensors<sup>12</sup>, supercapacitors<sup>13</sup> and high-performance batteries<sup>14</sup>.

Mesoporous materials are also gaining attention for biomedical applications such as drug or gene delivery15,16. Mesoporous silicas, in particular, can be synthesized in various shapes and sizes, are often biocompatible and spontaneously degrade in human tissues – a property that could be used to release drugs trapped in the silica. Moreover, the ability to accurately control the diameters of mesopores in silica is expected to provide tremendous advantages in biomedical applications, because the pore sizes directly affect the loading and release kinetics of drugs in delivery systems.

The main uses envisaged for mesoporous materials include as adsorbents in industrial processes for separating chemicals, and as catalysts in petrochemical refinery processes. Indeed, the original motivation for Kresge and colleagues' MCM-41 research was to synthesize catalytic materials for petroleum refining<sup>17</sup>. But although MCM-41 had sufficiently large pores for this purpose, its glass-like amorphous framework showed poor catalytic activity.

Ever since, enormous efforts have been made to synthesize mesoporous materials that contain crystalline, microporous, zeolite-like frameworks, which exhibit high catalytic performance. A breakthrough was made ten years ago, with the report of a specially designed surfactant molecule that enables the synthesis of such materials 18,19. The catalytic properties of the resulting mesoporous zeolites have not been fully explored for industrial processes, because the required surfactant is costly and not yet commercially available. However, I expect that mesoporous zeolites will trigger the next explosion of research in this field, by opening up many opportunities for catalytic applications.

Ryong Ryoo is at the Center for

Nanomaterials and Chemical Reactions. Institute for Basic Science, Daeieon 34141, South Korea, and in the Department of Chemistry, KAIST, Daeieon.

e-mail: rryoo@kaist.ac.kr

- Kresge, C. T., Leonowicz, M. E., Roth, W. J., Vartuli, J. C. & Beck, J. S. Nature 359, 710-712 (1992).
- Yanagisawa, T., Shimizu, T., Kuroda, K. & Kato, C. Bull. Chem Soc. Inn 63 988-992 (1990)
- Cao, L., Man, T. & Kruk, M. Chem. Mater. 21, 1144-1153 (2009)
- Stein, A. et al. Chem. Mater. 7, 304-313 (1995).
- Zhang, F. et al. J. Am. Chem. Soc. 127, 13508-13509 (2005).
- Attard, G. S. et al. Angew. Chem. 109, 1372-1374 (1997).
- Zhao, D. et al. Science 279, 548-552 (1998).

- 8. Rvoo, R., Joo, S. H. & Jun, S. J. Phys. Chem. B 103. 7743-7746 (1999).
- Joo, S. H. et al. Nature 412, 169-172 (2001).
- 10. Yang, H. & Zhao, D. J. Mater. Chem. 15, 1217-1231 (2005).
- 11. Liang, C., Li, Z. & Dai, S. Angew. Chem. Int. Edn 47, 3696-3717 (2008).
- 12. Nengqin, J., Wang, Z., Yang, G., Shen, H. & Zhu, L. Electrochem. Commun. 9, 233-238 (2007).
- 13. Li, W. et al. Carbon **45**, 1757–1763 (2007).
- 14. Ji, X., Lee, K. T. & Nazar, L. F. Nature Mater. 8, 500-506
- 15. Vallet-Regi, M., Balas, F. & Arcos, D. Angew. Chem. Int. Edn 46, 7548-7558 (2007).
- 16. Han, Y., Stucky, G. D. & Butler, A. J. Am. Chem. Soc. 121, 9897-9898 (1999).
- Kresge, C. T. & Roth, W. J. Chem. Soc. Rev. 42, 3663-3670 (2013).
- 18. Choi, M. et al. Nature **461**, 246–249 (2009).
- 19. Na. K. et al. Science 333, 328-332 (2011).

#### **Palaeontology**

## **Evolutionary insights from** Australopithecus

#### **Dean Falk**

In 1925, a *Nature* paper reported an African fossil of a previously unknown genus called Australopithecus. This finding revolutionized ideas about early human evolution after human ancestors and apes split on the evolutionary tree.

Australian-born Raymond Dart had barely started his job as chair of the anatomy department of the University of the Witwatersrand in Johannesburg, South Africa, when he made a momentous discovery. Using his wife's knitting needles, he painstakingly extracted a fossil (Fig. 1) from a chunk of rock found in Taungs (now known as Taung), South Africa. As he recalled<sup>1</sup>, "the rock parted ... What emerged was a baby's face, an infant with a full set of milk teeth ... I doubt if there was any parent prouder of his offspring than I was of my 'Taungs baby' on that Christmas of 1924." Better yet, the fossil fitted neatly with another type of fossil, called an endocast, formed from sediments accumulated inside the skull. The endocast reflects brain-surface details stamped on the braincase's inner walls. These fossils revealed a combination of apelike and human-like features never previously reported together.

Convinced that the specimen, called the Taung Child, represented an extinct link between humans and our ape ancestors, Dart dispatched a report<sup>2</sup> to *Nature* by mail boat. He probably felt some trepidation because several fellows of the Royal Society in London, who had mentored and taught with him, considered the human forerunner to be the British specimen known as Piltdown Man (which was later exposed as a hoax). Piltdown Man's human-sized brain and ape-like jaw contrasted with the Taung Child's ape-sized brain and

human-like jaw and teeth. In Dart's view, the Taung Child looked more primitive and older than the main existing candidates for the earliest ancestral human relative - Piltdown Man and Java Man (Homo erectus) from Indonesia. Dart therefore described the Taung Child as a 'man-ape' rather than an 'ape-man', like Java Man, and named the species Australopithecus africanus, which means southern ape from Africa.

Dart declared that humankind's cradle was not in Indonesia or Britain as his contemporaries thought, but was instead in Africa, as Charles Darwin had previously suggested<sup>3</sup>. The comfortable habitats favoured by African chimpanzees and gorillas in Dart's time were more than 3,200 kilometres north of where the Taung Child dwelled, and Dart suggested in his 1925 Nature paper that intense competition for limited resources in harsh southern African landscapes "furnished a laboratory such as was essential to this penultimate phase of human evolution". In the paper, he also reasoned that "enhanced cerebral powers possessed by this group ... made their existence possible in this untoward environment", attributing intelligence based on his interpretation of human-like brain convolutions at the back of the specimen's endocast.

When the paper appeared, the Taung Child and 32-year-old Dart became world famous overnight. Yet not everyone was receptive to new ideas about human evolution. Indeed,

#### 10 extraordinary papers

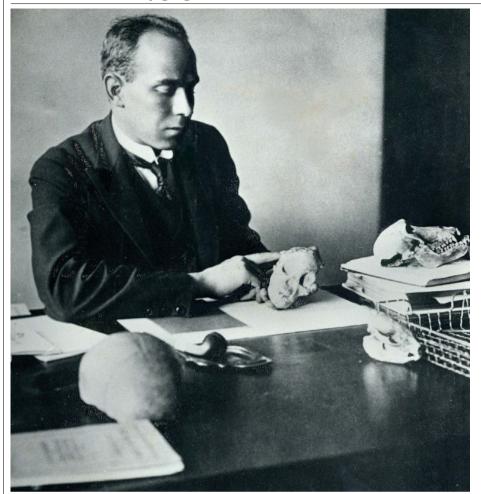


Figure 1 | Raymond Dart in 1925 holding the Australopithecus africanus fossil called the Taung Child.

five months later, a court case known as the Scopes monkey trial began in the United States to settle whether evolution could be taught in Tennessee schools. The immediate reaction to Dart's paper was mainly enthusiastic, but he soon became a target of 'you'll-burn-in-hell' letters from religious fundamentalists, and his former London colleagues published harsh criticisms of his research. Dart's main champion, the physician Robert Broom, remarked4: "It makes one rub one's eyes. Here was a man who had made one of the greatest discoveries in the world's history – a discovery that may yet rank in importance with Darwin's Origin of Species; and English culture treats him as if he had been a naughty schoolboy."

To answer his critics, Dart spent four years preparing a book<sup>5</sup> about the Taung Child. It provided voluminous extra details about the endocast, bones and teeth, and bolstered the argument that humans originated in Africa6. He submitted the book to the Royal Society, which declined to publish it. The pro-Piltdown fellows were probably behind this rejection7. Sadly, the book remains unpublished.

The most controversial aspect of Dart's paper, then and now, is his view that the back of the Taung Child's endocast is human-like. Some have argued that Dart misidentified a

skull imprint as a brain groove similar to a human one, a feature that is inconsistent with the Taung Child's otherwise ape-like brain<sup>8</sup>. Dart's 1925 Nature paper describes two endocast brain grooves, but his book identifies 14 further grooves, and describes 3 dispersed brain regions that look expanded in comparison with those of ape brains. If these findings had been published, they might have influenced the still-controversial debate about whether the human brain evolved in a piecemeal, mosaic fashion or in a more globally connected manner. Some mosaicists still cite Dart's 1925 *Nature* paper, but his unpublished book reveals his globalist viewpoint.

Dart's paper stated: "we may confidently anticipate many complementary discoveries concerning this period in our evolution." Indeed, thousands of specimens have been found that represent various Australopithecus species that lived in Africa during different time spans from more than 4 million to around 1 million years ago. The fossil Lucy is an example of one such species, called Australopithecus afarensis.

Subsequent work confirmed that Dart got most of the details right regarding his discovery. Australopithecus shared features of both living apes and humans, and they were bipedal

as he surmised because the skull opening that accommodates the spinal cord is positioned centrally at the base of the specimen's cranium. Dart correctly inferred9 that hominins originated in Africa, and that our genus Homo arose from Australopithecus. Happily, he lived long enough to see his initially iconoclastic ideas become widely accepted.

I cannot help but wonder what Dart would have thought about another notable discovery reported in *Nature*<sup>10</sup> – the 2004 identification of a species called Homo floresiensis (the most complete specimen is nicknamed the Hobbit) from remains in Indonesia dating to approximately 100,000-60,000 years ago. Like the Taung Child, the H. floresiensis specimens showed a combination of features never previously found in a fossil specimen. Homo floresiensis had ape-like, Australopithecus-like and human-like traits, as well as a tiny brain, leading some to suggest that this species might be a lineage descended from a previously unknown early hominin migration out of Africa11.

The parallels with Dart's discovery are remarkable. *Homo floresiensis* drew worldwide attention, but was also met with scorn from some scientists (who argued that the Hobbit represents an abnormal human). Homo floresiensis-like fossils dating to 700,000 years ago have since been reported12, and its legitimacy as a species is gaining 🕏 traction. It might be equally crucial for unravelling the evolution of early members of the human family tree outside Africa in the way that the Taung Child was essential for understanding the evolution of human ancestors in Africa. Only time will tell. One thing is certain, however; the more palaeoanthropology changes, the more palaeopolitics stays the same.

Dean Falk is in the Department of Anthropology, Florida State University, Tallahassee, Florida 32310, USA, and is also at the School for Advanced Research, Santa Fe, New Mexico.

e-mail: dfalk@fsu.edu

- Dart, R. A., with Craig, D. in Adventures with the Missing Link 10 (Harper, 1959)
- Dart, R. A. Nature 115, 195-199 (1925).
- Darwin, C. The Descent of Man, and Selection in Relation to Sex (Appleton, 1871)
- Broom, R. in Finding The Missing Link 27 (Watts, 1950). Dart, R. A. Australopithecus africanus: And His Place
- in Human Origins (unpublished manuscript, Univ. Witwatersrand Archives, 1929) 6. Falk, D. Am. J. Phys. Anthropol. 140 (Suppl. S49), 49-65
- (2009)
- Falk, D. The Fossil Chronicles: How Two Controversial Discoveries Changed Our View of Human Evolution (Univ. California Press, 2011).
- Keith, A. in New Discoveries Relating to the Antiquity of Man 84-85 (Norton, 1931). 9. Dart, R. A. J. Hum. Evol. 2, 417-427 (1973).
- 10. Brown, P. et al. Nature 431, 1055-1061 (2004).
- 11. Argue, D., Groves, C. P., Lee, M. S. Y. & Jungers, W. L. J. Hum. Evol. 107, 107-133 (2017).
- 12. van den Bergh, G. D. et al. Nature 534, 245-248 (2016).

#### **Astronomy**

### First exoplanet found around a Sun-like star

#### Eliza Kempton

In 1995, astronomers detected a blisteringly hot Jupiter-mass planet orbiting closer to its host star than Mercury is to the Sun. This discovery recast our thinking of how planets form and led to a new era of exoplanetary exploration.

Anyone over the age of 35 will remember growing up in a world in which only one planetary system was known - our own. We remember proudly reciting the names of the nine planets (eight before Pluto's discovery in 1930, and again today with its reclassification as a dwarf planet in 2006) and wondering what other planets might exist around the stars in the night sky. Contemplating life beyond the Solar System was relegated to science fiction. This all changed in 1995 when Mayor and Queloz<sup>1</sup> reported the detection of the first exoplanet around a Sun-like star.

The discovery of the gas-giant planet named 51 Pegasi b after its parent star, 51 Pegasi - came as a surprise. Gas-giant planets, such as Jupiter, are located in the outer parts of the Solar System. The prevailing theory was, and still is, that the formation of these planets requires icy building blocks that are available only in cold regions far away from stars. Yet Mayor and Queloz found 51 Pegasi b to be orbiting about ten times closer to its host star than Mercury is to the Sun (Fig. 1). One possible explanation is that the planet formed farther out and then migrated to its current location.

The gas-giant planet was not the first exoplanet to be discovered. However, the previous detections<sup>2,3</sup> were of even stranger objects orbiting pulsars - rapidly spinning neutron stars, which are the collapsed remnants of hot massive stars. The discovery of 51 Pegasi b was the first to substantiate the existence of planets around long-lived hydrogen-burning stars that resemble the Sun.

The bizarre character of a gas-giant planet orbiting so close to its parent star engendered considerable scepticism about the true nature of 51 Pegasi b. Mayor and Queloz detected the planet through minute back-and-forth motion of 51 Pegasi, which seemed to indicate that a planet-mass object was pulling on the star. But this stellar motion, sensed by frequency shifts in the spectra of light from 51 Pegasi, had other possible interpretations. A lively debate ensued in the literature about whether pulsations of the star might be masquerading as a planetary signature<sup>4,5</sup>.

This debate was put to rest in 1998 when the astronomer David F. Gray wrote a paper refuting his previous assertion that the stellar spectra were indicative of pulsations rather than a planet<sup>6</sup>. Further vindication came through the detection of planets similar to 51 Pegasi b, as other researchers combed their existing data for similarly unexpected planetary signals<sup>7</sup>. These highly irradiated giant planets have come to be known as hot Jupiters.

In the 24 years since the discovery of 51 Pegasi b, about 4,000 exoplanets have been identified (see go.nature.com/2jpcgtf). Other detection techniques have entered the scene, including the transit method, in which an exoplanet is revealed through the subtle dimming of its host star as the planet crosses the line of sight between Earth and the star. Hot Jupiters have continued to be discovered by the many exoplanet searches that are sensitive to large planets on close orbits. However, it is now known that such objects are intrinsically rare, orbiting only about 1% of Sun-like stars8.

By contrast, planets known as super-Earths and mini-Neptunes abound. Such objects. which inhabit the size and mass gap between the rocky and gas-giant planets of the Solar System, were also a surprise to planet hunters, but seem to be commonplace in our Galaxy. There is now good reason to think that the Milky Way contains more planets than it does stars9.

Mayor and Queloz's detection of 51 Pegasi b gave rise to a new field of astronomy. The ranks of exoplanet researchers have been steadily growing, by some counts now making up about one-quarter of the astronomy profession (see go.nature.com/32imc4i). Incipient subfields include the study of exoplanet demographics and the characterization of exoplanetary atmospheres.

This characterization has confirmed that hot Jupiters truly are gas-giant planets, but ones representing what our own Jupiter would look like if it were suddenly transported 100 times closer to the Sun. Amid the scorching-hot hydrogen-helium envelopes of these planets, astronomers have detected trace amounts of steam, carbon monoxide and metal vapours<sup>10-12</sup>. Such atmospheric studies could lead to the eventual characterization of exoplanets that resemble Earth.

The future of the exoplanet field is bright. In April 2018, NASA launched the Transiting Exoplanet Survey Satellite (TESS), a space telescope that is just beginning to fulfil its mission of finding small transiting planets around the brightest stars in the night sky. These planets will be ideally suited for follow-up using NASA's James Webb Space Telescope (JWST), once it launches, to measure their atmospheric properties and compositions. Following on the heels of JWST, the European Space Agency has selected the Atmospheric Remote-sensing Infrared Exoplanet Large-survey (ARIEL) space

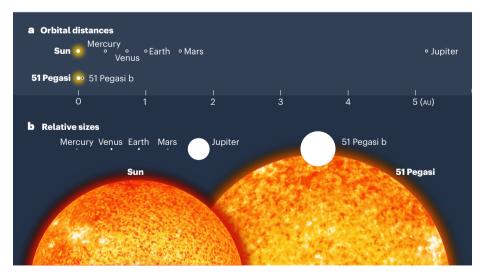


Figure 1 | The planetary systems of the Sun and of 51 Pegasi. a, In the Solar System, gas-giant planets, such as Jupiter, orbit far from the Sun. In 1995, Mayor and Queloz reported the discovery of 51 Pegasi b – a gasgiant planet that is much closer to its host star, 51 Pegasi, than Mercury is to the Sun. The orbital distances of the planets are given in astronomical units (1 AU is the average separation between Earth and the Sun). b, The sizes of all objects are shown approximately to scale.

#### 10 extraordinary papers

telescope to launch in 2028. ARIEL will be dedicated to characterizing the atmospheres of a wide sample of exoplanets.

These programmes are paving the way towards the ultimate goal of potentially detecting the signatures of life on an exoplanet. This goal could most optimistically be achievable in the next decade, but more realistically will require a new generation of space- and ground-based telescopes<sup>13</sup>. What is remarkable is that humans have gone from discovering the first exoplanets to legitimately plotting out the search for life on these worlds in just a quarter of a century.

**Eliza Kempton** is in the Department of Astronomy, University of Maryland,

College Park, Maryland 20742, USA. e-mail: ekempton@astro.umd.edu

- Mayor, M. & Queloz, D. Nature 378, 355-359 (1995)
- 2. Wolszczan, A. & Frail, D. A. Nature 355, 145-147 (1992).
- 3. Wolszczan, A. Science 264, 538-542 (1994).
- 4. Gray, D. F. Nature 385, 795-796 (1997)
- Hatzes, A. P., Cochran, W. D. & Johns-Krull, C. M. Astrophys. J. 478, 374–380 (1997).
- 6. Gray, D. F. Nature 391, 153-154 (1998).
- Butler, R. P., Marcy, G. W., Williams, E., Hauser, H. & Shirts, P. Astrophys. J. 474, L115–L118 (1997).
- 8. Howard, A. W. et al. Astrophys. J. Suppl. 201, 15 (2012).
- Batalha, N. M. Proc. Natl. Acad. Sci. USA 111, 12647–12654 (2014).
- 10. Kreidberg, L. et al. Astrophys. J. **793**, L27 (2014).
- Snellen, I. A. G., de Kok, R. J., de Mooij, E. J. W. & Albrecht, S. Nature 465, 1049–1051 (2010).
- 12. Hoeijmakers, H. J. et al. Nature 560, 453-455 (2018).
- National Academies of Sciences, Engineering, and Medicine. Exoplanet Science Strategy (National Academies. 2018).

#### **Cell biology**

# Cell identity reprogrammed

#### Samantha A. Morris

The discovery that cell differentiation can be reversed challenged theories of how cell identity is determined, laying the foundations for modern methods of reprogramming cell identity and promising new regenerative therapies.

All cells of an organism derive from a single cell. As development progresses, cells become increasingly specialized to perform defined functions, a commitment that is accompanied by a restriction in the range of potential fates of those cells. In the late nineteenth century. a predominant thought was that, when they differentiate, cells retain only those pieces of heritable information required to maintain cell-type identity and function<sup>1</sup>. This led to the theory that differentiation is an irreversible process (Fig. 1a). John Gurdon's seminal paper in *Nature* on nuclear reprogramming of cell identity, with Tom Elsdale and Michael Fischberg<sup>2</sup>, provided a remarkable challenge to this dogma, and formed the basis for today's cell-reprogramming field.

Gurdon and colleagues' 1958 paper was preceded by the work of Robert Briggs and Thomas King<sup>3</sup>. To investigate the developmental potential of differentiating cells, Briggs and King used a method called nuclear transfer, in which the nucleus is removed from one cell (in this case, an egg) and replaced with an intact nucleus from a different cell. Briggs and King's experiments were a technical feat that had previously been accomplished only in single-celled organisms<sup>4</sup>.

Using this method in the more-complex Northern leopard frog (*Rana pipiens*), they

were able to produce normal, swimming tadpoles by replacing egg-cell nuclei with nuclei from blastomeres — cells that are made through the splitting of a fertilized egg cell during early development<sup>3</sup>. However, the transfer of nuclei from *R. pipiens* cells at more-advanced stages of differentiation — from when the hollow ball of blastomeres differentiates into a multilayered structure called a gastrula, onwards — did not support the development of normal frogs<sup>5</sup> (Fig. 1b).

#### "Since this paper appeared, biologists have developed the ability to reprogram cell identity by several routes."

Thus, Briggs and King's results demonstrated that the nuclei in blastomeres are not irreversibly changed with differentiation. However, they also indicated that, as development progresses, the potential of transplanted nuclei to support normal development decreases – suggesting that cell differentiation might be irreversible and might involve irreversible genetic changes. Thus, Briggs and King concluded<sup>5</sup> that the nuclei of cells in the latestage gastrula have an "intrinsic restriction

in potentiality for differentiation".

In 1958, Gurdon, Elsdale and Fischberg addressed the questions surrounding the potential of differentiated cells using a different species of frog, *Xenopus laevis* (the African clawed frog). In contrast to the *Rana* species, whose availability is seasonally restricted, *X. laevis* is available year round and rapidly reaches sexual maturity<sup>2</sup>. In the authors' experiments, donor nuclei from cells at various developmental stages, from early blastomeres to cells from tadpoles just before hatching, were transferred into *Xenopus* egg cells.

The donor nuclei were derived from a mutant stock in which each cell contained only one nucleolus (an organelle inside the nucleus) instead of the usual two. This approach provided a useful visual marker to confirm that the resulting animals obtained from nuclear transfer were indeed derived from the transferred nucleus, and not from existing material in the egg. These experiments demonstrated that normal tadpoles could be obtained from cells at stages of development up to pre-hatching tadpole stages (Fig. 1c) — much later than the developmental stage of the cells that Briggs and King had used.

Many of the tadpoles that developed from cells containing transferred nuclei underwent normal metamorphosis into frogs, which seemed to be sexually mature. The authors noted that the lone frog derived from the most-differentiated cell nucleus was "accidentally killed shortly before metamorphosis". A subsequent report<sup>6</sup> was free of such misadventure; it described the derivation of fertile adult frogs from the transplanted nuclei of fully differentiated cells collected from the intestines of feeding tadpoles.

Gurdon and colleagues thus demonstrated, unlike Briggs and King, that differentiated nuclei could support successful development. Despite this discordance, both groups agreed that the advance of a nucleus through differentiation was accompanied by a reduction in its ability to support normal development. On the basis of their findings that some differentiated nuclei could support normal development (albeit with a relatively limited frequency of success), Gurdon and colleagues concluded that the differentiated cell state is not a result of irreversible genomic changes. Rather, the nuclei of differentiated cells retain the capacity to orchestrate the development of a fully functioning organism.

Almost 40 years after these amphibian experiments, transfer of the nucleus of an adult mammary epithelial cell was used to generate a cloned mammal: Dolly the sheep<sup>7</sup>. The first mouse to be cloned using nuclear transfer from adult cells, Cumulina, was reported shortly afterwards<sup>8</sup>. To prove beyond doubt that cloned animals could be produced using nuclei from fully differentiated cells (and had not previously been derived from contaminant

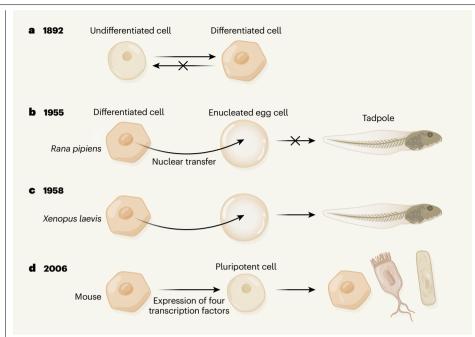


Figure 1 | Key milestones in understanding the potential of differentiated cells. a, In 1892, Weismann proposed that, as cells in a developing embryo differentiate, they retain only those genes required to maintain cell-type identity, rendering differentiation an irreversible process<sup>1</sup>. b, Studying the Northern leopard frog (Rana pipiens), Briggs and King reported<sup>5</sup> in 1955 that nuclei from differentiated cells that were transferred into an egg cell from which the nucleus had been removed (an enucleated egg cell) could not support normal development, in line with Weismann's thinking. c, In their 1958 Nature paper<sup>2</sup>, Gurdon, Elsdale and Fischberg challenged the notion that development is irreversible, reporting that nuclei derived from differentiated cells of the African clawed frog (Xenopus laevis) could, in fact, support normal development, d, In 2006, Takahashi and Yamanaka<sup>13</sup> identified a core set of four transcription factors that reset differentiated mouse cells to a pluripotent state, capable of giving rise to any cell types in the body.

stem cells that had broader potential), mice were derived using the nuclei of mature B cells and T cells<sup>9</sup>. During maturation, the genomes of both of these types of immune cell undergo DNA rearrangements, which were detected in the clones.

Together, this rich history of nuclear transfer revealed that cell differentiation can be reversed, resetting cell identity to the earliest embryonic stages. This pioneering work formed the foundations for the reprogramming field, which has the core goal of manipulating cell identity to produce any desired cell type.

In the 1980s, early work in reprogramming revealed that it is possible not only to reset cell identity to the blank slate of early embryonic development, but also to switch a cell's identity altogether. For example, one study10 showed that fusion of a mouse muscle cell with a human amniocyte (a fetal cell that floats in the amniotic fluid) to produce a cell with both a human and a mouse nucleus resulted in the rapid expression of human muscle-specific genes. This showed that factors produced in a differentiated cell (in this case, the mouse muscle cell) can induce the expression of genes that are repressed in another type of differentiated cell (in this case, the human amniocyte). Together with the nuclear-transfer studies, these pivotal experiments established that factors produced in egg cells and differentiated cells are able to direct cell fate by regulating gene expression.

A key moment came in 1987, when a single factor capable of reprogramming cell identity was identified; the expression of a protein called MyoD (a transcription factor) was shown to convert fibroblast cells into contracting muscle cells11. Gurdon was somewhat pessimistic about the prospect that cell reprogramming could be quickly achieved using a defined set of factors, stating in 2006, "Looking far ahead, it may become possible to convert cells of an adult to an embryonic state without needing to use eggs"12. However, just a few months later, Kazutoshi Takahashi and Shinya Yamanaka reported that differentiated cells could be reset to a pluripotent state – that is, a state in which they could differentiate into multiple types of cell – through the expression of only four transcription factors<sup>13</sup> (Fig. 1d). In 2012, Gurdon and Yamanaka were awarded the Nobel Prize in Physiology or Medicine for their work.

Since Gurdon and colleagues' paper demonstrating that developmental potential can be reinstated in differentiated cells, cell biologists have developed the ability to reprogram cell identity by several routes. For example, we can use transcription-factor-mediated reprogramming to return cells to an embryonic state<sup>13</sup> and subsequently direct their differentiation to desired identities by mimicking normal developmental processes. Alternatively, embryonic states can be altogether avoided by expressing specific factors to directly convert a differentiated cell type to another cell identity14-16. Such strategies offer the potential to produce patient-derived cells for modelling diseases in vitro17.

Moreover, cell reprogramming forms the basis of various proposed regenerative therapies, including the generation of cells that line the retina at the back of the eye to treat a disorder called age-related macular degeneration<sup>18</sup>, a major cause of vision loss.

Gurdon and colleagues' 1950s conclusions that the developmental clock can be reset challenged the long-standing theory at that time that cell differentiation is an irreversible process. Their work now represents a cornerstone of current reprogramming technologies that aim to deliver a range of cell types for disease modelling and regenerative therapies.

Samantha A. Morris is in the Departments of Developmental Biology and of Genetics, and in the Center of Regenerative Medicine, Washington University School of Medicine in St Louis, St Louis, Missouri 63110, USA. e-mail: s.morris@wustl.edu

- Weismann, A. The Germ-Plasm: A Theory of Heredity (transl. Parker, W. N. & Ronnfeldt, H.) (Scott, 1893).
- Gurdon, J. B., Elsdale, T. R. & Fischberg, M. Nature 182. 64-65 (1958).
- Briggs, R. & King, T. J. Proc. Natl Acad. Sci. USA 38, 455-463 (1952).
- DiBerardino, M. A. & Hoffner, N. J. Results Probl. Cell Differ, 11, 53-64 (1980).
- King, T. J. & Briggs, R. Proc. Natl Acad. Sci. USA 41, 321-325 (1955)
- Gurdon, J. B. J. Embryol. Exp. Morphol. 10, 622-640
- 7. Campbell, K. H., McWhir, J., Ritchie, W. A. & Wilmut, I. Nature 380, 64-66 (1996). Wakayama, T., Perry, A. C., Zuccotti, M., Johnson, K. R. &
- Yanagimachi, R. Nature 394, 369-374 (1998).
- 9. Hochedlinger, K. & Jaenisch, R. Nature 415, 1035-1038
- 10. Blau, H. M., Chiu, C. P. & Webster, C. Cell 32, 1171-1180 (1983).
- 11. Davis, R. L., Weintraub, H. & Lassar, A. B. Cell 51, 987-1000 (1987)
- 12. Gurdon, J. B. Annu. Rev. Cell Dev. Biol. 22, 1-22 (2006).
- 13. Takahashi, K. & Yamanaka, S. Cell 126, 663-676 (2006).
- 14. Cohen, D. E. & Melton, D. Nature Rev. Genet. 12, 243-252
- 15. Morris, S. A. & Daley, G. Q. Cell Res. 23, 33-48 (2013).
- 16. Vierbuchen, T. & Wernig, M. Nature Biotechnol. 29 892-907 (2011).
- 17. Passier, R., Orlova, V. & Mummery, C. Cell Stem Cell 18, 309-321 (2016)
- 18. Mandai, M. et al. N. Engl. J. Med. 376, 1038-1046 (2017).

# The discovery of the Antarctic ozone hole

#### Susan Solomon

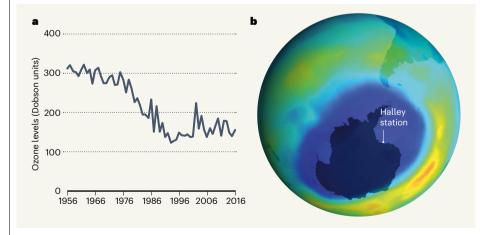
The unexpected discovery of a hole in the atmospheric ozone layer over the Antarctic revolutionized science — and helped to establish one of the most successful global environmental policies of the twentieth century.

In 1985, Joe Farman, Brian Gardiner and Jonathan Shanklin reported<sup>1</sup> unanticipated and large decreases in stratospheric ozone levels over the Antarctic stations of Halley and Faraday. Their data showed that, after about 20 years of fairly steady values, ozone levels began dropping in the austral spring months around the late 1970s (Fig. 1). By 1984, the stratospheric ozone layer over Halley in October was only about two-thirds as thick as that seen in earlier decades – a phenomenon that became known as the Antarctic ozone hole. Farman et al. boldly suggested a link to human use of compounds called chlorofluorocarbons (CFCs), often used in aerosol cans and cooling devices such as fridges. Their findings transformed the fields of atmospheric science and chemical kinetics, and led to global changes in environmental policy.

The stability of the stratospheric ozone layer has attracted the interest of scientists, the public and policymakers for more than 50 years because this layer protects life on Earth's surface from biologically damaging ultraviolet

radiation. The potential for pollutants known as nitrogen oxides to deplete global ozone prompted much research² on the influence of aviation on the ozone layer³. A study⁴ in 1974 suggested that chlorine monoxide (ClO) produced from CFCs might similarly deplete ozone. By the early 1980s, the best projections from stratospheric models indicated that continuing production of CFCs at then-current amounts risked the destruction of only about 2–4% of the ozone layer by the end of the twenty-first century³. There was no suggestion that ozone at polar latitudes would be especially sensitive.

The expected depletion was relatively small and far in the future, but posed serious threats, including increased incidence of skin cancers and ecological damage. International policymakers therefore concluded that a cautious ozone-protection strategy was needed, and, in March 1985, the United Nations Vienna Convention for the Protection of the Ozone Layer was signed. It called for more ozone research, but contained no legally



**Figure 1** | **Ozone over Antarctica.** a,  $\ln 1985$ , Farman et al. reported that stratospheric ozone levels over the Halley and Faraday stations in Antarctica during the austral spring had declined greatly from previously steady values. The graph shows the Halley times series, extended to 2016. **b**, Subsequent satellite monitoring revealed that the area of ozone depletion – the ozone hole – extended over a vast region. This map shows a satellite ozone map for 10 September 2000, when ozone depletion was close to its maximum: blue indicates low ozone levels; red, high levels. The position of the Halley station is indicated.

binding goals for CFC reductions5.

Farman and colleagues' report of a loss of one-third of the springtime ozone layer over Antarctica was published a few months later. The paper's strengths were the authors' careful analysis of the seasonal character of the change, and the fact that changes were detected using two different instruments. The authors suggested that Antarctica's extremely cold temperatures during winter and spring made the region "uniquely sensitive to growth of inorganic chlorine" produced in the atmosphere from CFCs, although the chemical mechanism they proposed was incorrect. The careers of hundreds of scientists and dozens of diplomats worldwide were abruptly transformed by this single paper.

At that time, the atmospheric chemistry of the Antarctic was *terra incognita*. Measurements needed to be made both at ground level and from aircraft to understand whether CFCs had a role in producing the ozone hole. Scientists were energized and excited to attack the challenge.

I was fortunate to be among a group of scientists who went to the US station at McMurdo in 1986, where the first Antarctic measurements of CIO (ref. 6) and of another CFC-derived ozone-depleting compound, chlorine dioxide (OCIO) (ref. 7), were obtained. These compounds were roughly 100-fold more abundant than elsewhere. The 'smoking gun' for the role of CFCs in ozone depletion came from aircraft measurements taken in 1987. They revealed a dramatic enhancement in CIO levels (comparable to those at McMurdo) and a co-located decrease in ozone concentrations as the plane flew south from Chile into the Antarctic.

These independently obtained data sets indicated that the Antarctic was indeed uniquely sensitive to chlorine compounds<sup>9</sup>, as Farman *et al.* had suggested. Unusual changes in atmospheric abundances of related chemicals were also measured<sup>10</sup>. Moreover, satellite monitoring confirmed that depletion extended over a vast region (typically up to about 20 million square kilometres; see ref. 10, for example).

The response of policymakers to Farman and colleagues' paper was initially cool. In my view, this was because they did not want to upset the apple cart of the delicate diplomacy embarked on with the Vienna convention until it was clear that the science was correct. Nevertheless, they argued that precautionary principles were part of the convention, and – even as the research planes were flying from Chile – signed the 1987 Montreal Protocol on Substances that Deplete the Ozone Layer. This was an agreement to freeze production and consumption of ozone-depleting substances at then-current rates, and to meet over time to consider whether to decrease production.

But the signing of global environmental

agreements is only a ceremonial first step; they must subsequently be ratified and strengthened over time5. I believe that Farman and colleagues' paper led to the remarkably fast ratification of the protocol in 1989, and to later amendments (beginning with the London Amendment in 1990) that included ever-tightening restrictions on the global production and consumption of ozone-depleting substances.

So why was the ozone hole not seen in computational simulations of the stratosphere? It turned out that the models lacked a key ingredient: by considering only gas-phase atmospheric chemistry, they overlooked the activation of ozone-destroying chlorine species that occurs on and within polar stratospheric cloud particles at extremely low temperatures<sup>11,12</sup>. The discovery of the missing ingredient drew physical chemists in increasing numbers to study the surface chemistry involved13. Previously unknown gasphase reactions associated with ozone depletion were also identified, particularly those involving a ClO dimer (see ref. 10, for example). Laboratory and field studies were carried out, and microphysical models were developed (see ref. 14, for example), to determine what polar stratospheric clouds are made of: ice, nitric acid hydrates or supercooled liquids. The answer was that they could be all three, depending on temperature and the histories of the sampled air parcels.

Ground-based and airborne missions to understand Arctic ozone chemistry15 were also inspired by Farman and colleagues' paper and related studies. It emerged that ozone loss in the Arctic is generally much less severe than in the Antarctic, broadly because temperatures in the region are warmer as a result of meteorological differences between the two regions. The coupling of chlorine-containing species with bromine-containing ones was found to be a key ingredient in polar ozone depletion, especially in the Arctic<sup>16</sup>.

Atmospheric modelling also progressed to simulate the newly discovered processes, evolving from two dimensions (latitudealtitude) to three (latitude-altitude-longitude), to better represent global stratospheric temperatures, winds and circulation<sup>17</sup>. Dynamical studies have shown that the ozone hole influences Antarctic winds and temperatures not just in the stratosphere, but also in the underlying troposphere, and there is evidence for climate connections at other latitudes18. Modern global climate models therefore include increasingly detailed representations of stratospheric chemistry and dynamics. The ozone hole has thus inspired a new generation of scientists to probe climate-chemistry interactions, forging connections between previously separate disciplines.

The Montreal Protocol led to global CFC production and consumption phase-outs by 2010, and now the Antarctic ozone hole is slowly healing<sup>10</sup>. The protocol thus prevented the ozone layer from collapsing<sup>19</sup> and is a signature success story for global environmental policy. Because CFCs have atmospheric lifetimes of 50 years or more, the atmosphere will not fully recover until after 2050, even in the absence of further emissions.

However, recent work<sup>20</sup> provides strong evidence of the continuing production and release of one type of CFC (trichlorofluoromethane). The source is not large enough to reverse the healing of the ozone hole, but it is slowing recovery and shows that there is still a need for scrutiny in this field. Research into, and policy to protect, the stratosphere will thus continue to be inspired by Farman and colleagues' research – and will probably do so until the ozone hole finally closes.

Susan Solomon is in the Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. e-mail: solos@mit.edu

- Farman, J. C., Gardner, B. G. & Shanklin, J. D. Nature 315. 207-210 (1985).
- Crutzen P. L.O. L.R. Meteorol. Soc. 96, 320-325 (1970).
- National Research Council, Causes and Effects of Changes in Stratospheric Ozone: Update 1983 (Natl Acad. Press, 1984).
- Molina, M. J. & Rowland, F. S. Nature 249, 810-812 (1974).
- Benedick, R. A. Ozone Diplomacy: New Directions in Safeguarding the Planet (Harvard Univ. Press, 1998).
- de Zafra, R. L. et al. Nature 328, 408-411 (1987).
- Solomon, S., Mount, G. H., Sanders, R. W. & Schmeltekopf, A. L. J. Geophys. Res. Atmos. 92, 8329-8338 (1987).
- Anderson, J. G. et al. J. Geophys. Res. Atmos. 94, 11480-11520 (1989).
- Solomon, S. Nature 347, 347-354 (1990).
- 10. World Meteorological Organization. Scientic Assessment of Ozone Depletion: 2018 - Report No. 58 (WMO. 2018).
- 11. Solomon, S., Garcia, R. R., Rowland, F. S. & Wuebbles, D. J. Nature 321 755-758 (1986)
- 12. Tolbert, M. A., Rossi, M. J., Malhotra, R. & Golden, D. M. Science 238, 1258-1260 (1987).
- 13. Ravishankara, A. R. & Hanson, D. R. J. Geophys. Res. Atmos. 101, 3885-3890 (1996).
- 14. Peter, T. & Groos, J.-U. in Stratospheric Ozone Depletion and Climate Change (ed. Muller, R.) Ch. 4, 108-144 (R. Soc. Chem., 2011).
- 15. Pyle, J. A. et al. Geophys. Res. Lett. 21, 1191-1194 (1994).
- 16. Frieler, K. et al. Geophys. Res. Lett. 33, L10812 (2006).
- 17. Eyring, V. et al. Atmos. Chem. Phys. 10, 9451-9472 (2010).
- 18. Thompson, D. W. J. et al. Nature Geosci. 4, 741-749 (2011).
- 19. Newman, P. A. Atmos. Chem. Phys. 9, 2113-2128 (2009).
- 20. Montzka, S. A. et al. Nature **557**, 413-417 (2018).

#### **Immunology**

# The advent and rise of monoclonal antibodies

#### Klaus Rajewsky

A 1975 Nature paper reported how cell lines could be made that produce an antibody of known specificity. This discovery led to major biological insights and clinical successes in treating autoimmunity and cancer.

In their 1975 *Nature* paper<sup>1</sup>, the immunologists Georges Köhler and César Milstein described the production of monoclonal antibodies of predetermined specificity, each made by a continuously growing cell line that had been generated by the fusion of an antibody-producing cell from an immunized mouse with an immortal cancer cell specialized for antibody secretion. Hearing from César about this work before it was published, on the way to an obscure meeting in San Remo in Italy, I knew immediately that our research field had reached a turning point.

Antibodies were discovered in 1890 by the physiologist Emil von Behring and the microbiologist Shibasaburo Kitasato as protective antitoxins in the blood of animals exposed to diphtheria or tetanus toxin<sup>2</sup>. Ever since, antibodies have been a major research subject, given their key role in adaptive immunity (specific immune responses against, for example, invading disease-causing agents) and

their wide range of specificities, essentially covering the universe of chemical structures. This had stood out from early on as a major genetic puzzle. How can our limited genome encode a seemingly limitless repertoire of specificities? And in medical (and industrial) practice, antibodies have been used ever since their discovery as the basis for serum therapy (the treatment of infectious diseases using blood serum from immunized animals), as diagnostic tools to monitor infectious disease, and in innumerable other contexts.

But antibodies specific for any given molecule (called an antigen in the context of an antibody response) came, with a few notable exceptions, as mixtures of antibodies, produced by thousands of antibody-producing cells in an immunized animal or infected person. Each of these cells produced an antibody of its own kind, so that 'antibody specificity' usually referred to the properties of antibody populations rather than those

### 10 extraordinary papers

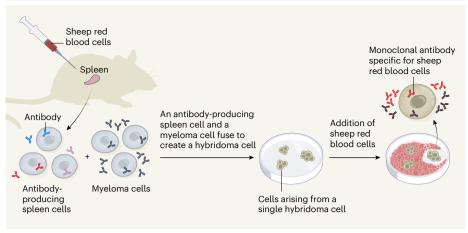


Figure 1 | The production of monoclonal antibodies. Köhler and Milstein's 1975 Nature paper¹ solved the problem of how to generate clones of continuously dividing cells that make antibodies of a known specificity. The ability to generate such monoclonal antibodies revolutionized antibody research and paved the way to clinical advances. The authors injected mice with sheep red blood cells and isolated spleen cells, including those that produce antibodies. Different antibody colours indicate antibodies specific for different molecules (antigens), and produced by different cells. The authors had the idea of fusing antibody-producing spleen cells of limited lifespan with myeloma cells – immortal cancerous immune cells secreting antibodies of unknown specificity. Spleen cells that had been activated upon antigen recognition fused preferentially with the myeloma cells, generating hybrid cells called hybridomas. Unlike unfused cells, the hybridoma cells could grow on the selective agar plates used, and formed colonies of identical cells. Hybridomas that secreted antibodies specific for sheep red blood cells were identified by their ability to destroy such cells when added to the agar, generating a clearance (plaque). These original hybridoma cells made two types of antibody, one that recognized sheep red blood cells and another of unknown specificity.

of individual antibodies. The inability to produce molecularly defined, homogeneous antibodies of predetermined specificity was a major hurdle that needed to be overcome.

This changed overnight with Köhler and Milstein's paper. Köhler had joined Milstein's group at the MRC Laboratory of Molecular Biology in Cambridge, UK, as a postdoc, to study the mechanism of somatic mutation that operates in antibody diversification. The plan was to use mouse myeloma cells for this purpose. These are tumour cells originating from antibody-secreting immune cells. The cancer immunologist Michael Potter at the National Cancer Institute in Bethesda, Maryland, had shown years before that myelomas could be induced in a particular mouse strain by the injection of mineral oil<sup>3</sup>. The Milstein team was propagating and fusing to each other cells obtained from cell lines derived from various such tumours. However, the myeloma antibodies were ill-defined in terms of specificity. Could one perhaps fuse antibody-producing cells from immunized mice to myeloma cells, to produce continuously dividing cells that make antibodies specific for the immunizing antigen? To detect such fused cells, an approach offered itself which Köhler had become acquainted with during his PhD at the Basel Institute for Immunology in Switzerland and that had been developed by the institute's director, Niels Jerne<sup>4</sup>. This was a simple technique in which cells secreting antibodies in response to, and specific for, sheep red blood cells (SRBCs) can be identified by the formation of a clearance (called a plaque) in SRBC-containing agar plates.

With this, the stage was set for the Köhler-Milstein experiment (Fig. 1). Large numbers of plaque-forming hybrid cells secreting anti-SRBC antibodies appeared when spleen cells from SRBC-immunized mice were fused with myeloma cells. The fused cells had acquired expression of a single type of anti-SRBC antibody from a spleen cell and preserved the immortality and high rate of antibody secretion of the myeloma fusion partner. Myeloma and spleen cells were unable to multiply under the chosen experimental conditions. and the myeloma cells apparently preferred antigen-activated spleen cells over others for fusion, a prerequisite for the striking success of the experiment.

The fused cells could be cloned and propagated indefinitely as what were later termed hybridomas, producing unlimited amounts of monoclonal antibodies. The first-generation hybridomas secreted two types of antibody: the desired one, plus an antibody of unknown specificity originating from the myeloma fusion partner. But this two-antibody problem was soon solved through the isolation of myeloma lines that had lost antibody expression<sup>5,6</sup>.

Antibodies against any desired antigen could now be generated, investigated and used as homogeneous molecular entities. In 1984, Köhler and Milstein won the Lasker Award together with Potter, and that same year Köhler, Milstein and Jerne were awarded the Nobel Prize in Physiology or Medicine.

The impact of the Köhler-Milstein paper on biomedical and, specifically, immunological research was dramatic, propelled by scientific developments that occurred around the time the paper appeared. Thus, it became clear shortly afterwards that the variable and constant regions of antibodies are encoded by separate gene segments. Antibody diversity arises when somatic recombination joins gene segments together, and when a subsequent process called somatic hypermutation operates, during the course of the antibody response, on the recombined gene segments encoding antibody variable regions. Together. these mechanisms generate a vast repertoire of antibody specificities, as well as distinct classes of antibody, which mediate their various roles (effector functions) through their differing constant regions.

These insights were accompanied by the explosive development of new molecular and genetic tools that allowed the isolation and manipulation of antibody genes in multiple ways. Together with the hybridoma technology, they fuelled a rapidly growing and still expanding field of investigation, in which basic research on antibody diversification and effector function goes hand-in-hand with the production and engineering of monoclonal antibodies for diagnostic and therapeutic purposes.

In the early days, the production of monoclonal antibodies was entirely based on hybridoma technology and used for two main purposes: to study the somatic evolution of the antibody repertoire and the molecular basis of antibody specificity; and to generate reagents that bind to specific proteins or other molecules expressed by cells of the body or by pathogens. In both cases, completely new insights and technical advances resulted. Thus, affinity maturation of antibodies (the increase of antibody affinity during the course of an antibody response) began to be understood at the molecular level. And the technique of fluorescence-activated cell sorting was revolutionized by monoclonal antibodies, allowing the separation of different cell types at an unprecedented level of specificity and resolution. Recent highlights in this area include approaches allowing gene-expression profiling of single cells that have been characterized by the expression of large arrays of surface-marker proteins through cocktails of DNA-tagged, 'barcoded' monoclonal antibodies<sup>7</sup>.

In medicine, monoclonal antibodies have an ever-increasing role and have generated a multibillion-dollar market, which is expected to grow substantially in the future. In addition to their impact on medical diagnosis, the therapeutic application of antibodies has led to spectacular successes in the treatment of autoimmune diseases and cancer. The 2018 Nobel Prize in Physiology or Medicine was awarded for the "discovery of cancer therapy by [antibody-mediated] inhibition of negative

immune regulation". As often happens in biology, both the mechanisms and the efficient induction of the inhibitory processes underlying this type of immunotherapy are still unclear, with ongoing research providing challenges and new perspectives that are driving the development of monoclonal antibodies against additional targets.

Monoclonal antibodies are also being developed to control infectious diseases – following the concept of protective antibodies that goes back to von Behring and Kitasato. Prevalent diseases such as malaria, influenza and AIDS call for the development of what are termed broadly neutralizing monoclonal antibodies, which, applied individually or in cocktails, might provide broad protection<sup>8</sup>.

Intensive work in this direction has yielded promising results, including engineering antibody specificity through the substitution of variable domains by ligand-binding domains from non-antibody receptors9. Yet the immune system itself uses similar tricks10 and, by and large, antibody design is still unable to outdo it in terms of generating and selecting antibody specificities<sup>11</sup>. Nevertheless, the manifold modern molecular, cellular and genetic approaches to selecting and engineering antibodies have had, and continue to have. a tremendous impact on the field, whether by producing partly or fully human antibodies of different classes, making bi-specific or toxin-conjugated antibodies for specific

therapeutic purposes, or incorporating antibody variable regions into chimaeric antigen receptors on T cells for use in an anticancer treatment called CAR-T cell therapy.

Monoclonal antibodies are nowadays often generated by isolating or transforming antibody-producing cells taken directly from immunized animals or patients, and transplanting the antibody-encoding genes of these cells into suitable producer cell lines, rather than using hybridoma technology<sup>12-14</sup>. But they started their spectacular career in 1975, secreted by hybridoma cells in Köhler and Milstein's SRBC-containing agar plates.

Klaus Rajewsky is at the Max-Delbrück-Center for Molecular Medicine in the Helmholtz Association, 13125 Berlin, Germany. e-mail: klaus.rajewsky@mdc-berlin.de

- Köhler, G. & Milstein, C. Nature 256, 495-497 (1975).
- Behring, E. & Kitasato, S. Dtsch Med. Wochenschr. 49, 1113-1114 (1890).
- Potter, M. & Boyce, C. R. Nature 193, 1086-1087 (1962).
- Jerne, N. K. & Nordin, A. A. Science 140, 405 (1963).
- Kearney, J. F., Radbruch, A., Liesegang, B. & Rajewsky, K. J. Immunol. 123, 1548-1550 (1979).
- Galfrè, G, & Milstein, C. Methods Enzymol. 73, 3-46 (1981).
- Todorovic, V. Nature Methods 14, 1028-1029 (2017).
- Walker, L. M. et al. Science 326, 285-289 (2009).
- Gardner, M. R. et al. Nature 519, 87-91 (2015).
- 10. Tan, J. et al. Nature 529, 105-109 (2016)
- Verkoczy, L., Alt, F. W. & Tian, M. Immunol. Rev. 275, 89-107 (2017).
- 12. Scheid, J. F. et al. Nature 458, 636-640 (2009).
- 13. Steinitz, M., Klein, G., Koskimies, S. & Mäkelä, O. Nature 269, 420-422 (1977)
- 14. Traggiai, E. et al. Nature Med. 10, 871-875 (2004).

#### Materials science

# The nano-revolution spawned by carbon

#### Pulickel M. Ajayan

In 1985, scientists reported the discovery of the cage-like carbon molecule  $C_{60}$ . The finding paved the way for materials such as graphene and carbon nanotubes, and was a landmark in the emergence of nanotechnology.

The history of the carbon molecule C<sub>60</sub> highlights the fact that discoveries do not happen in a predefined sequence. C<sub>60</sub>, carbon nanotubes and graphene (single layers of graphite) are essentially members of the same family: all are nanoscale structures that consist of carbon atoms arranged in a periodic crystal lattice. Graphite has been known for a few hundred years, and individual layers of the material could be separated easily. However, the identification of C<sub>60</sub> by Kroto et al.<sup>1</sup> did not occur until 1985. This, in turn, led to the discovery of graphene nearly two decades later<sup>2</sup>. Both of these breakthroughs led to

Nobel prizes, in chemistry for  $C_{60}$  (1996) and in physics for graphene (2010).

The discovery of C<sub>60</sub> occurred on the campus of Rice University in Houston, Texas. Eiji Osawa, a Japanese theoretical chemist, had predicted<sup>3</sup> the stable structure of a 60-atom carbon molecule in 1970, but this finding did not come to the attention of the mainstream scientific community. Experimental results from mass spectrometry were also beginning to emerge, showing the stability of 60-atom carbon clusters. However, no one made the connection that these clusters would have the structure that Osawa had predicted. It

## 150 years ago

Aphorisms by Goethe — the opening article of the first issue of Nature. 4 November 1869.

Nature! We are surrounded and embraced by her: powerless to separate ourselves from her, and powerless to penetrate beyond her. Without asking, or warning, she snatches us up into her circling dance. and whirls us on until we are tired, and drop from her arms. She is ever shaping new forms: what is, has never yet been; what has been, comes not again. Everything is new, and yet nought but the old ... So far Goethe.

When my friend, the Editor of NATURE, asked me to write an opening article for his first number, there came into my mind this wonderful rhapsody on "Nature", which has been a delight to me from my youth up. It seemed to me that no more fitting preface could be put before a Journal, which aims to mirror the progress of that fashioning by Nature of a picture of herself, in the mind of man, which we call the progress of Science

[In a letter to Chancellor von Müller] Goethe says, that about the date of this composition of "Nature" he was chiefly occupied with comparative anatomy; and in 1786, gave himself incredible trouble to get other people to take an interest in his discovery, that man has a intermaxillary bone. After that he went on to the metamorphosis of plants; and to the theory of the skull: and, at length. had the pleasure of his work being taken up by German naturalists. The letter ends thus:-"If we consider the high achievements by which all the phenomena of Nature have been gradually linked together in the human mind ... we shall, not without a smile ... rejoice in the progress of fifty years."...

When another half-century has passed, curious readers of the back numbers of NATURE will probably look on our best, "not without a smile;" and, it may be, that long after the theories of the philosophers whose achievements are recorded in these pages, are obsolete, the vision of the poet will remain as a truthful and efficient symbol of the wonder and the mystery of Nature.

#### T. H. Huxley



### 10 extraordinary papers

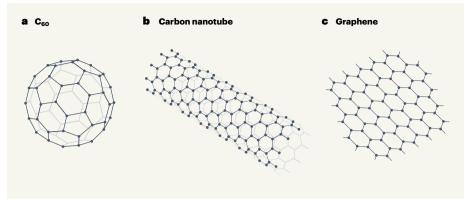


Figure 1 | Three major nanoscale carbon structures discovered in the past 35 years. a,  $\ln 1985$ , Kroto  $et al.^1$  reported the discovery of the molecule  $C_{60}$ . It has a cage-like structure that consists of 12 pentagonal and 20 hexagonal faces. b, Following Kroto and colleagues' work, carbon nanotubes were first produced  $^{10}$  in 1991. A carbon nanotube can be thought of as a 2D hexagonal lattice of carbon atoms that is rolled up to form a hollow cylinder. c,  $\ln 2004$ , scientists reported the isolation of graphene  $^2$  – a single layer of carbon atoms in a 2D hexagonal lattice.

was against this backdrop that the visit of the British chemist Harry Kroto to the laboratories of Rice scientists Richard Smalley and Robert Curl proved significant.

Kroto was an expert in molecular spectroscopy and had an interest in the molecules that exist in interstellar space. He proposed a simple mechanism for the formation of the small carbon-chain molecules that had been observed in interstellar gas clouds, and suggested that this idea could be tested using Smalley's experimental apparatus. Smalley, Curl and their students were making many different atomic clusters, such as those of silicon, through ablation — the removal of material from the surface of a target — and were analysing the masses of these clusters in detail. After some delay, Kroto's proposal was accepted and he journeyed to Houston.

In previous work by other groups<sup>4</sup>, a peak corresponding to  $C_{60}$  was somewhat prominent in mass spectra. During the experiments at Rice to test the mechanism of carbon-chain formation, it became clear that the  $C_{60}$  peak could be made extremely strong under certain conditions. However, the structure of the  $C_{60}$  molecule was the main puzzle that needed to be solved. The team accomplished this task, and published the first report in 1985.

The structure of  $C_{60}$  turned out to be a beauty (Fig. 1a). It looked exactly like the classic design of a football (soccer ball). More precisely, the structure is about 0.7 nanometres across and is a truncated icosahedron — a polyhedron that has 12 pentagonal and 20 hexagonal faces. This highly symmetric, cage-like shape was first described by Archimedes, and the rules that guide the topology of polyhedra were first developed by Descartes.

When applied to polyhedra that are made of only pentagons and hexagons, these rules imply that every such closed structure can contain any number of hexagons but must have exactly 12 pentagons. Heptagons can also be introduced, producing negative curvature (saddle-shaped surfaces), but the topological effect of a heptagon is cancelled by that of a pentagon. In the mid-eighteenth century, the Swiss mathematician Leonhard Euler had proposed a formula for these geometric rules, which were now profoundly manifested at the nanoscale in  $C_{60}$ . Larger closed carbon cages (such as  $C_{70}$  and  $C_{82}$ ) also exist, and can be formed by simply adding more hexagons to the cage.

The family of  $C_{60}$  and larger molecules have come to be known as the fullerenes, after the US architect Buckminster Fuller. Fuller had become famous for designing stable domes

# "The structure of C<sub>60</sub> turned out to be a beauty. It looked exactly like the classic design of a soccer ball."

and buildings<sup>6</sup> that have shapes similar to that of  $C_{60}$ . The correspondence was striking, although the scale differed by a factor of about 10 billion. So it was that the  $C_{60}$  family got its name (its members could well have been called soccerenes).

Kroto and colleagues' fullerene discovery took other scientists by surprise. Initially, there were quite a few sceptics; many thought that  $C_{60}$  was flat rather than cage-like. However, this perception changed after work by the German chemist Wolfgang Krätschmer, the US chemist Donald Huffman and their students. In 1990, these researchers succeeded in isolating  $C_{60}$  molecules from carbon soot in bulk, thereby making the substance available for large-scale experiments.

The fullerene discovery immediately had two major consequences. First, fullerenes were used to synthesize a large variety of unconventional materials. For example, endohedrals<sup>8</sup>

(fullerenes that enclose metal atoms), fullerene-assembled solids and superconducting fullerene materials were produced and characterized with excitement. Fullerenes were seen as a distinctive, stable molecular system and as an ideal building block for making unprecedented materials. They were also touted as a new allotrope (structural form) of carbon that deviated from the familiar graphite and diamond.

Second, the discovery provided the impetus to seek other carbon allotropes – particularly nanoscale materials. The most substantial result from this search was the synthesis and development of carbon nanotubes (Fig. 1b) by the Japanese physicist Sumio lijima<sup>10</sup> and colleagues<sup>11</sup> in the early 1990s. Carbon nanotubes showed that the electronic structures of carbon layers could be tuned by structural nanoscale engineering, suggesting possible uses in electronics and other applications.

Over the following two decades, a rush of research activities, publications and patents would make fullerenes and carbon nanotubes the poster children of nanotechnology. It was also during this period, in 2004, that the Russian physicists Andrei Geim and Konstantin Novoselov isolated graphene<sup>2</sup> (Fig. 1c). Graphene was the first example of a truly stable 2D material and revealed the physics associated with such 2D systems.

It has been nearly 35 years since Kroto and colleagues' fullerene paper was published. In spite of all the potential that fullerenes promised, these molecules have not led to any major applications, barring a few encouraging ideas in solar cells and biochemistry. However, the work paved the way for many innovations in nanomaterials that will ultimately find uses in nanotechnology. The fullerene discovery and what followed show the ingenuity of the human mind in solving a nanoscale puzzle. Fullerenes also provide a curious case in which an architect's name was dragged into a major scientific discovery. Buckminster Fuller would probably not have minded.

**Pulickel M. Ajayan** is in the Department of Materials Science and NanoEngineering, Rice University, Houston, Texas 77006, USA. e-mail: ajayan@rice.edu

- Kroto, H. W., Heath, J. R., O'Brien, S. C., Curl, R. F. & Smalley, R. E. Nature 318, 162–163 (1985).
- 2. Novoselov, K. S. et al. Science **306**, 666–669 (2004).
- 3. Osawa, E. *Kagaku* **25**, 854-863 (1970).
- Rohlfing, E. A., Cox, D. M. & Kaldor, A. J. Phys. Chem. 81, 3322–3330 (1984).
- Euler, L. Novi Comm. Acad. Sci. Imp. Petropol. 4, 109–140 (1758).
- Sieden, L. S. Buckminster Fuller's Universe: His Life and Work (Basic, 2000).
- Krätschmer, W., Lamb, L. D., Fostiropoulos, K. & Huffman, D. R. Nature 347, 354–358 (1990).
- 8. Chai, Y. et al. J. Phys. Chem. 95, 7564-7568 (1991).
- Hebard, A. F. et al. Nature 350, 600-601 (1991).
   Iijima, S. Nature 354, 56-58 (1991).
- 11. Ebbesen, T. W. & Ajayan, P. M. *Nature* **358**, 220–222 (1992).

# Correspondence

### Darwin brokered others' publications

Charles Darwin was not just an occasional contributor who used Nature to share his own findings and discussions (see Y. Liu Nature 574, 36; 2019). He also gave voice to naturalists around the world, at a time when the journal was not easily accessible to the international scientific community.

Roughly half of Darwin's contributions to Nature were transcripts (with due credit) of reports and findings sent to him. The writers of the letters were from Brazil, the United States. Peru and Poland. His most frequent correspondent was Iohann Friedrich Theodor (Fritz) Müller, a German scientist who emigrated to southern Brazil in 1852. Müller, like several of Darwin's interlocutors, tested and observed facts described in On the Origin of Species (1859).

Müller's support for the theory of evolution was expressed in his book Für Darwin (1864). Darwin considered it of such importance that he himself sponsored its translation into English, published the year of Nature's launch, 1869, under the title Facts and Arguments for Darwin. This initiated a 17-year friendship between the two naturalists, documented by intensive correspondence. Between 1874 until the year before his death in 1882, Darwin transcribed seven of the scientific reports he received from Müller and submitted them to Nature.

Antonio C. Marques, Klaus Hartfelder University of São Paulo, Brazil. marques@ib.usp.br

### Scientism and the abuse of science

The philosopher Friedrich Hayek originally popularized the term 'scientism' in his 1979 book The Counter-Revolution of Science as a synonym for pseudoscience. The word later came to represent the expansion of science into domains where it really has nothing to say, such as evolution into atheism. Nathaniel Comfort now positions 'scientism' as the abuse of science in ways that obscure today's concerns for equity, inclusion and diversity (see N. Comfort Nature 574, 167-170; 2019).

Comfort condemns this version of scientism, in which practices and policies endorsed by scientists have had adverse consequences for vulnerable groups in society - although he is careful not to brand the scientists involved as malicious or ignorant. The implication is that history should help to ensure that such scientism will not happen again.

In my view, it is a misuse of history to oversee the future. What counts as good and bad in scientific practice or in science-based policies can be understood only in retrospect, because our judgement depends on witnessing the consequences. As we move forward in history, those judgements will change. It follows that the moral character of any action is indeterminate at the time it happens. Science itself is a quantum phenomenon – and 'scientism' is its observer effect.

Steve Fuller University of Warwick, Coventry CV4 7AL, UK. s.w.fuller@warwick.ac.uk

### **Engage egg donors** in editing debate

We argue that egg donors should be more involved in discussions on the ethical aspects of human germline gene editing (see Nature 574, 465-466 (2019) and E.S. Lander et al. Nature 567. 165-168: 2019).

Experimental data from large numbers of human embryos could be necessary to refine and improve germline gene editing, as well as to evaluate the technique's safety and efficacy. Moreover, studies involving the creation of embryos seem preferred for testing for specific mutations and to reduce mosaicism (H. Ma et al. Nature 548, 413-419; 2017). This means that oocytes will have to be procured from large numbers of women.

Oocyte harvesting exposes the donors to serious shortand long-term health risks, raising questions about the ethical acceptability of experiments that require this procedure. Although donors are often compensated for the inconvenience, the practice prompts concerns about undue inducement – particularly for financially vulnerable women. The ethical issues are exacerbated because it is by no means certain that clinical applications of germline gene editing will eventually be permitted.

Above and beyond the physical risks, these wider ethical and policy issues should be made clear to potential donors so that they can make an informed choice and have a chance to be properly engaged in the debate.

Emilia Niemiec, Heidi Carmen Howard Uppsala University, Sweden. heidi.howard@crb.uu.se

### How will we fund open-access fees?

The international Plan S research-funder consortium cOAlition S proposes that institutional libraries should transition from subscription to 'pure publish' deals with openaccess journals by 2024 (see Nature 572, 586; 2019). However, the coalition represents just 16 European funding agencies and 3 international charity foundations. Many other European funders are not in a position to pay open-access publication fees on behalf of their researchers.

For example, Denmark's 14,000 private foundations that currently support half of the country's research are stretched to the limit. Their researchers will therefore have no choice but to pay the bill out of their own research grants, which are already under intense pressure from spiralling costs.

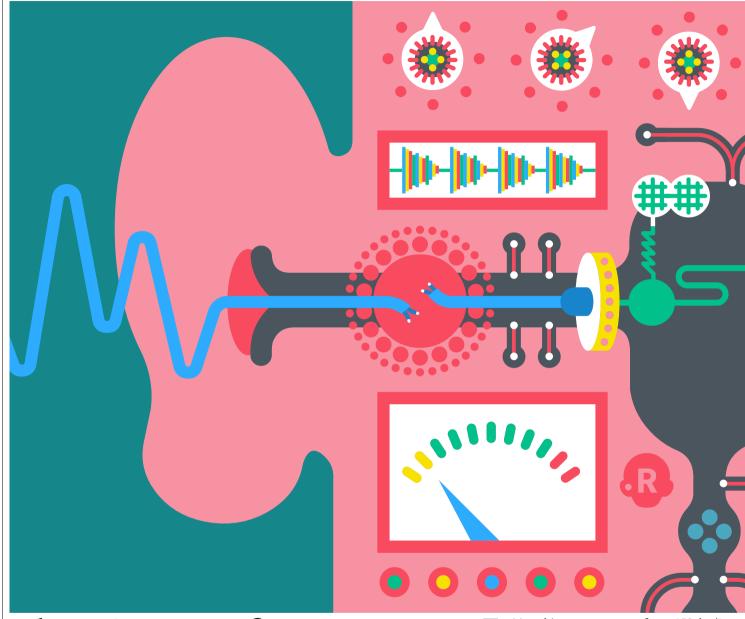
Remedial action is urgently needed if publication and knowledge flow are not to be skewed towards the wealthiest countries and universities. For example, national or European Union funds could be established to help cashstrapped researchers cover their publishing costs.

Christian Sonne, Rune Dietz, Aage K. O. Alstrup Aarhus University, Denmark. cs@bios.au.dk

#### **HOW TO SUBMIT**

Correspondence may be submitted to correspondence@ nature.com after consulting the author guidelines and section policies at go.nature.com/ cmchno.

# **Prize essays**



# The winners of our youngwriter essay competition

The contest for 18–25-year-olds received more than 660 entries from 68 countries.

n May this year, as part of our 150th anniversary, *Nature* asked readers aged between 18 and 25 to enter an essay competition. The task was to tell us, in no more than 1,000 words, what scientific advance they would most like to see in their lifetimes, and why it mattered to them.

The response was phenomenal: we received 661 entries. Some entrants hoped that science would make their lifetimes much longer than they can currently expect. Many looked forward to work that will end climate change. Others wanted to see advances in neuro-degenerative disease, our understanding of human history, crop growth, space exploration, medical technologies, water resources

ILLUSTRATIONS BY IAN KA



or superfoods. The standard of writing was impressive, and the scope of ideas inspiring.

The winner is a compelling essay by Yasmin Ali, a PhD student at the University of Nottingham, UK. Ali submitted a thought-provoking piece on Beethoven, her brother's hearing loss and the science she hopes will one day cure it. It stood out to the judges as a reminder of why many scientists do research: to make the world better tomorrow than it is today.

All essays were judged by a group of Nature editors. The top ten submissions were then ranked by three members of a separate judging panel: Magdalena Skipper, editor-in-chief of Nature; Faith Osier, an immunologist and researcher at the KEMRI-Wellcome Trust Research Programme in Kilifi, Kenya; and Jess Wade, a physicist at Imperial College London. All submissions were kept anonymous throughout the process.

We also selected two runners-up. Physicist Robert Schittko at Harvard University in Cambridge, Massachusetts, proposes that nuclear fusion could offer a solution to the climate crisis, in a piece that effortlessly mixes grand ambition with gentle humour. And chemist Matthew Zajac at the University of Chicago in Illinois wrote a powerful personal account of why he wants to see advances in the field of same-sex reproduction.

The results show that today's young scientists have a wealth of ideas, talent and conviction that research can transform their world. We look forward to seeing what they do next.

# Beethoven's dream

The composer wished for a cure for his hearing loss. Soon, research could make it a reality for my twin brother - and millions more. By Yasmin Ali

n 1802, under a June Sun, a 31-year-old Beethoven paced through the countryside around Vienna. Rays of sunshine pierced through the trees, the hard soil crunched beneath his feet and birds conducted their own orchestra. But Beethoven didn't marvel at these details; he was preoccupied by thoughts of suicide. Some years earlier, he had started to lose his hearing, and although it wasn't yet severe, he was still struggling immensely with his condition. Living with hearing loss made his life a "wretched existence" that drove him into despair, he wrote. He still persevered with his work, and went on to create timeless music. But he found little joy in the process.

I observed a similar struggle at first hand, as my twin brother Islam, when we were 18 years old, started to lose his hearing. I noticed changes in his personality, too. He was always the outgoing troublemaker, but became quiet and withdrawn. Because hearing loss isn't visible, I didn't know what he was going through, which also made it difficult for me to be there

Today, 466 million people worldwide have disabling hearing loss, and over 900 million are expected to have it by 2050, according to the World Health Organization. Its impact is often underestimated compared with other disabilities, but people with hearing loss constantly  $experience\,communication\,difficulties\,in\,their$ everyday lives. They often mishear speech and find it very difficult to follow conversations. These miscommunications can lead to individuals feeling isolated as they struggle to take part in social interactions, ultimately leading

them to withdraw from society. As Helen Keller once wrote: "Blindness cuts us off from things, but deafness cuts us off from people."

To this day, there is still no cure for sensorineural hearing loss (the most common type, and the one Beethoven had). We have advanced technological devices that amplify sound, such as hearing aids and cochlear implants, but these still don't restore hearing. In my and my brother's lifetimes, I'd like to see research make that possible.

Sensorineural hearing loss occurs as a result of damage to the inner ear organ, called the

### "If it works, such a scientific advance could transform hearing health care."

cochlea, which has intricate sound-sensing hair cells that are responsible for hearing. In humans and other mammals, any damage to hair cells is irreversible. Other animals, such as birds, fish, amphibians and reptiles, can spontaneously regenerate their cochlear hair cells, meaning that any hearing loss they develop is only temporary.

Scientists have been studying the regeneration process of hair cells in non-mammals, and have identified various genes and proteins that have central roles. These can be targeted to stimulate support cells in the cochlea to in turn create more hair cells and replace those that died.

Some of these cell therapies have been

### **Essay competition**

successful in restoring the hearing of mice and guinea pigs: a breakthrough! These advances have led to the development of more therapies, and one such therapy is now being tested for the first time in humans. The REGAIN clinical trial (REgeneration of inner ear hair cells with GAmma-secretase INhibitors), an international collaboration led by researchers at University College London, is testing a molecule called  $\gamma$ -secretase inhibitor that could potentially restore hearing by encouraging supporting cells to transform into new hair cells themselves.

If it works, such a scientific advance could transform hearing health care as we know it. My own research investigates the impact hearing loss has on people's mental well-being. Many people share Beethoven's despair when they realize that their hearing can't be restored. Hope is an essential element for good mental health.

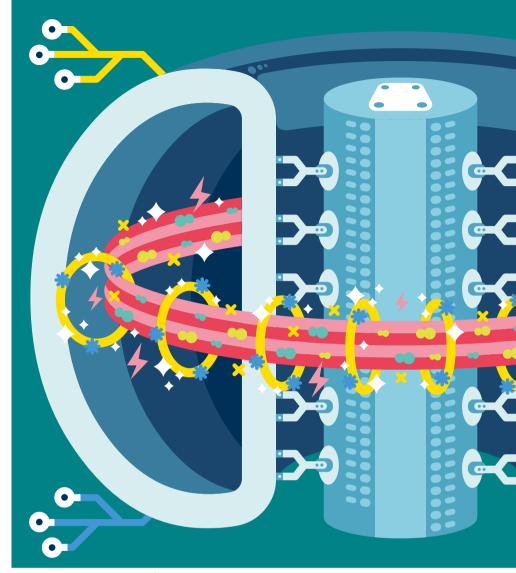
Other members of the deaf community see themselves as a cultural minority, rather than as a disabled group to be 'cured'. My and other scientists' research aims to help those who feel disadvantaged by deafness and want to be able to hear.

Islam and I come from interracial parents, so we look very different. I have white, freckled skin, and his is olive (he gets perfect suntans, and I turn into a tomato). I have blue eyes, and his are hazelnut. I have normal hearing, whilst he has severe hearing loss. He and I have shared the many chapters of our lives, and when things became difficult as his hearing declined, what helped us cope was being able to make sense of it all together. Communication, self-expression, hearing and being heard (even through sign language) are basic human needs. I hope that when I voice support to my brother in the future, that he'll be able to hear it, receive it and not feel alone.

When Beethoven lost his hearing, he secluded himself from society — but one thing that gave him strength was the hope that his hearing could be regained one day. But each medical remedy he attempted failed. In 1802, he wrote: "But, think that for six years now I have been hopelessly afflicted, made worse by senseless physicians, from year to year deceived with hopes of improvement, finally compelled to face the prospect of a lasting malady (whose cure will take years or perhaps be impossible)."

Beethoven's dream of regaining his hearing did not come true for him, but through the scientific advance of regeneration of hair cells, it could become a reality 217 years after his June walk. On his deathbed, it is said that Beethoven's last words were "I shall hear in heaven!" Luckily for us, those facing hearing difficulties could soon be able to hear on Earth.

**Yasmin Ali** is a PhD student studying mental health and hearing loss at the University of Nottingham, UK.



# Power play

# Nuclear-fusion power plants could be part of a solution to the climate crisis. By Robert Schittko

f primary sources can be believed, I conducted my first experiment with a high-power energy source at the tender age of one.

It was New Year's Eve 1995, and I had somehow gained possession of two silver objects I now know were screws, when my wandering gaze was captured by a snake-like item emerging from a wall. At its end, which I was about to learn was the head of an extension cord, there were two tiny openings, whose black interiors stood out daringly against the white backdrop of a piece of plastic. Utterly unaware of the cautionary tale I was about to write, I abandoned all hesitation. I took one last

breath, homed in on my target, and shoved the two silver objects into the two little holes, thus producing the first – but, fortunately, not the last – negative result of my newfound career.

Twenty-four years later, my parents and I have fully recovered from our respective shocks, I am still playing around with hazardous equipment – currently as a physicist at Harvard University in Cambridge, Massachusetts – and the mishandling of energy sources on a far larger scale is now threatening not just my existence but that of tens of thousands of species worldwide. Unlike toddler-me, today, we cannot plead ignorance. Even if global warming is kept to 1.5 °C above



pre-industrial levels, the Intergovernmental Panel on Climate Change (IPCC) has warned, "climate-related risks to health, livelihoods, food security, water supply, human security. and economic growth" will increase. Warming beyond that point to just 2.0 °C will further harm hundreds of millions of people in vulnerable areas worldwide, the IPCC estimates. Yet the emission levels countries have volunteered to aim for following the Paris agreement will warm Earth by approximately 3.0 °C over the next 80 years alone, and it seems that even these goals will not be met.

This failure of the global political establishment to adequately address climate change has prompted a hunger for some sort of transformative breakthrough, either of the political or of the technological kind.

Our best hope for the former - already expressed in a global wave of climate activism - might be an unprecedented political movement which dramatically ups the pressure to act more determinedly in the face of a crisis.

Our best hope for the latter is called nuclear fusion.

Nuclear fusion is a process by which pairs of light atomic nuclei unite while releasing enormous amounts of energy. It is the mechanism that powers the Sun and other stars, and a principle that researchers have long hoped to harness to build nuclear-fusion power plants. In theory, such plants could be fuelled with sustainably sourced hydrogen isotopes for thousands of years, while being safer than nuclear-fission plants and producing zero long-lived nuclear waste. Unfortunately, they also come with a catch: building them is incredibly hard.

This is because nuclear fusion on Earth requires temperatures in the order of tens of millions of degrees Celsius, at which the fusion fuel behaves as a riotous plasma. The difficulty in governing the behaviour of this plasma is the key reason why nuclear-fusion power plants do not exist today, despite over sixty years of extensive research. Nevertheless, those years have resulted in many valuable insights, and a clean-energy future thanks to nuclear fusion seems more realistic today than ever before.

The most ambitious nuclear-fusion project to date, ITER, is currently being constructed in southern France with the explicit goal of pushing past break-even, a so-far elusive point of operation at which the output power of the fusion process exceeds the power invested to maintain the plasma. Helped by dozens of other labs around the world, ITER, which is scheduled to start full operation in 2035, will also test several auxiliary technologies that a working fusion plant would ultimately require, all while separate research into competing types of fusion reactors continues elsewhere and breakthroughs such as deep learning advance the field (J. Kates-Harbeck et al. Nature 568. 526-531: 2019). With all this in mind, I'm hopeful that working

nuclear-fusion plants will be built well before the end of the century, and that fusion energy will help substantially in limiting the impact of the climate crisis.

Irrespective of that crisis, there are plenty of other reasons to be excited about nuclear fusion. As a physicist, I am humbled by the idea of taming a plasma that is several times hotter than the Sun's core. As a researcher, I am amazed by the complexity that a nuclear-fusion power plant would require in every aspect of its ultimate design. And as a writer, I marvel at the prospect of mimicking the stars, instead of merely looking up to them.

### "As a physicist, I am humbled by the idea of taming a plasma that is several times hotter than the Sun's core."

But it is as a human, thinking of other humans, that I feel a breakthrough in controlled fusion could rise above all else. After all, the human cost of climate change, of rising seas and rising temperatures, of more frequent droughts and extreme weather events, will ultimately have to be paid. And it will be paid first and foremost by those who have the least; by the poor and the less privileged, who can be faulted for the crisis they will be caught up in no more than a one-year-old boy can be faulted for electrocuting himself.

Nuclear-fusion power plants, more so than any other technology, could prove a uniquely powerful tool to diminish that cost.

That's why I hope to see them in my lifetime.

Robert Schittko is an experimental physicist at Harvard University, Cambridge, Massachusetts.

# Reproduction, rethought

### Same-sex partners should one day be able to raise a biological descendant together. By Matthew Zajac

ne afternoon as a second-year undergraduate, I called my parents from my dormitory. To them it was a routine call home, but to me it was a conversation long overdue. I'd rehearsed with my closest friends exactly how to start; my words needed to strike with confidence but should mitigate shock. Like protecting them from a grenade I'd thrown at them.

'So ... I actually do have some romance in my life. With a boy."

I practised answers to typical questions parents ask after their child comes out as gay: "Are you sure?", "Why haven't you told us?", "Didn't you like a girl once?" But those questions never came, and I wasn't prepared for the one my mom did ask: "What about kids?" Whether out of sympathy for my aspirations to raise children or because of her plans of pampering grandchildren, my mother quickly recognized that my ability to start a family could be jeopardized by



my sexuality. And she wasn't wrong; 74% of American adults are parents, but only 35% of lesbian, gay, bisexual and transgender (LGBT) adults are parents even though 51% express the desire to have children, according to a 2013 survey. As of 2015, two-thirds of minors living with same-sex couples come from a previous relationship. But this is changing. With homosexuality becoming more accepted in parts of the world, people are recognizing their sexual identity earlier and might be less likely to enter a different-sex marriage. As such, fewer same-sex couples are raising children, but those children are more likely to be born during a same-sex relationship.

This trend is partly due to increased opportunities for same-sex couples to parent, by adoption and other means. *In vitro* fertilization (IVF) and surrogacy offer partial genetic relatedness for same-sex female and male couples, respectively. However, neither of these options delivers full genetic relatedness. Although no evidence suggests that genetic relatedness is necessary or sufficient for parenthood, surveys of biologically infertile different-sex couples show its significance. In 2017, one study found that more than 97% of respondents would prefer having a genetically related child (S. Hendriks *et al. Hum. Reprod.* 32, 2076–2087; 2017).

Now, as a graduate student performing

research in chemical biology at the University of Chicago, Illinois, I think a lot about the intersection between my sexuality and my scientific interests. Genome-editing techniques are currently transforming our capacity to study fundamental biology. But, more importantly for me, they have offered a glimmer of hope that I could one day raise a biological descendant with my partner.

The road to same-sex human reproduction is one that many think is impossible to traverse. Aside from ethical and sociopolitical roadblocks, there are fundamental biological issues.

Parthenogenesis, or reproduction from an egg cell without fertilization, occurs naturally in birds and sharks. But mammalian reproduction is complicated by genomic 'imprinting', in which some genes are modified or shut down in either sperm or eggs while their opposite numbers are expressed – like the two halves of a zipper coming together. Seeking to address this, researchers have derived 'imprint-free' stem cells. A 2018 report in Cell Stem Cell described the use of CRISPR to delete imprinted regions from mouse genomes – removing the teeth from the biological zipper (Z.-K. Li et al. Cell Stem Cell 23, 665-676; 2018). Use of this technique with eggs from female mice produced living pups that grew to be healthy, fertile adults.

However, pups produced using the technique with sperm from male mice did not survive to adulthood. While a significant achievement, many see the low success rate of birth (14% with embryos from two mothers, 2.5% with embryos from two fathers) as proof that mammals are limited to sexual reproduction. However, the technique offers optimism that same-sex human reproduction may be possible with a better understanding of imprinting, among other advances.

The development of same-sex reproduction technology might in 2019 be a scientific fantasy, and its use would be controversial. But IVF and same-sex marriage would have been just as unthinkable in 1869, when *Nature* launched from a foundation of academic liberalism and bold science. The disruptive innovation of same-sex reproduction would simply continue this endeavour and provide children to capable parents, as long as it is investigated enough to eliminate risks, made financially accessible and regulated responsibly.

As for me, I aspire to give my parents a grandchild by any plausible means when my partner and I are ready. But to raise a child genetically related to me and my partner? That's a dream I'll always have.

**Matthew Zajac** is a chemical biologist at the University of Chicago, Illinois, USA.

# **News & views**

#### **Neurodegeneration**

# Selective clearance of mutant huntingtin protein

#### Huda Y. Zoghbi

Compounds have been found that reduce levels of the harmful protein present in Huntington's disease, without affecting the normal version. The compounds interact with the mutated protein and the cell's protein-clearance machinery. See p.203

Several neurodegenerative diseases involve the slow accumulation of a misfolded protein in neurons over many years. The proteins involved in these diseases might differ, but the result is similar – eventually, the neurons die from the build-up of toxic misfolded proteins. Scientists have long been searching for ways to reduce the levels of the disease-driving proteins without also clearing their wild-type counterparts, which typically have myriad crucial functions. On page 203, Li et al. show that this can be accomplished using compounds that interact specifically with both the misfolded part of the protein and the neuron's protein-clearance machinery.

Li and colleagues chose to focus on Huntington's disease, which is caused by an abnormally long stretch of glutamine amino-acid residues in the huntingtin (HTT) protein. This expanded polyglutamine tract causes HTT to misfold. Affected individuals typically carry one copy (one allele) of the HTT gene that encodes the mutant protein, and one allele that encodes a protein with the normal-length glutamine tract.

Cells are able to degrade the mutant huntingtin (mHTT) through autophagy<sup>2</sup> – a clearancemechanismthatinvolvesengulfment of proteins by a vesicle called the autophagosome. Li et al. hypothesized that compounds that bind to both the mutant polyglutamine tract and the protein LC3B, which resides in the autophagosome, would lead to engulfment and enhanced clearance of mHTT (Fig. 1). But no such compounds had been reported. The authors therefore conducted small-molecule screens to identify candidate compounds, and used wild-type HTT in a counter-screen to rule out compounds that bind to the normal version

Li and colleagues initially identified

two candidates, dubbed 1005 and 8F20. These compounds had been shown<sup>3,4</sup> to inhibit, respectively, the activity of the cancer-associated protein c-Raf and kinesin spindle protein (KSP), which has a key role in the cell cycle. The team found that 1005 and 8F20 were able to clear mHTT independently of their effects on these other proteins.

The researchers showed that the regions of 1005 and 8F20 that interacted with mHTT and LC3B in the screen shared structural similarities. Next, they screened for compounds

that shared these structural properties but were structurally distinct from other c-Raf and KSP inhibitors (a compound that acts on mHTT without altering these proteins would be desirable for clinical treatment). This led them to discover two more compounds, AN1 and AN2, that link mHTT to LC3B and thereby selectively reduce levels of mHTT.

Li and colleagues validated their exciting discovery by showing that the four compounds reduced levels of the full-length mHTT protein (not just the protein fragment used in the screen). The compounds lowered levels of mHTT both in vitro - in mouse neurons and neurons derived from the biopsied skin cells of people with Huntington's disease - and in vivo, in mouse and fly models of the disease.

A key strength of the compounds identified by Li and co-workers is that they leave levels of wild-type HTT unchanged. This is crucial because HTT has multiple neuronal functions, both during embryonic development and after birth<sup>5</sup>. Existing mHTT-lowering strategies typically affect both HTT alleles<sup>6,7</sup>, which is not ideal. Equally, the compounds found by Li et al. did not affect other proteins that contain polyglutamine tracts of variable, but not disease-causing, length. These proteins often have many roles in the brain.

One question that naturally arises is whether

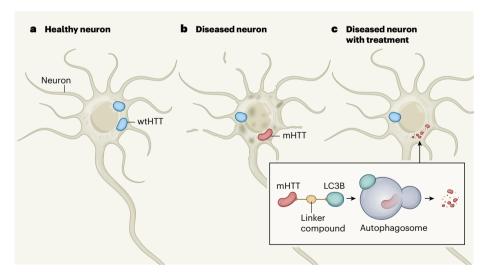


Figure 1 | Lowering levels of mutant huntingtin protein. a, Healthy neurons typically carry two copies of the gene that encodes the wild-type version of huntingtin protein (wtHTT). Only two proteins are shown, for  $simplicity, although \, many \, are \, produced \, from \, each \, gene \, copy. \, \boldsymbol{b}, \, Hunting ton's \, disease \, involves \, the \, expansion \, disease \, involves \, the \, expansion \, disease \, involves \, disease \, disea$ of a tract of glutamine amino-acid residues in one copy of HTT protein, producing a mutant version (mHTT) that accumulates in neurons, causing them to shrink and eventually die. Any strategy to decrease levels of mHTT in these cells must not affect wtHTT, which has key functions in the brain. c, Li et al. have identified four linker compounds that fulfil this role. Treatment with these compounds inhibits neuronal degeneration  $in \ various \ models \ of \ Hunting ton's \ disease. \ The \ compounds \ bind \ to \ both \ mHTT \ and \ the \ protein \ LC3B-a \ key$ component of a protein-clearance pathway called autophagy. This enables selective engulfment of mHTT by a vesicle called the autophagosome, leading to the mutant protein's degradation.

#### **News & views**

treating cells with the compounds led to enhanced autophagic clearance of proteins other than mHTT. Li et al. assessed the levels of the repertoire of proteins in the cortices of mice that carried an mHtt allele. They found changes in the abundance of a small percentage of proteins in mice treated with the compounds, compared with untreated animals. What remains unclear is whether the levels of some proteins decreased because mHTT levels were diminished, or because of autophagy. Modest changes in protein-expression level (in the 20-30% range for some wild-type proteins) can cause neurological deficits<sup>8</sup>, so pinpointing any off-target effects of the compounds will be a crucial next step. Even effects that initially seem inconsequential might build up over the course of long-term therapy, becoming as problematic decades later as the original toxic protein.

Despite these concerns, the authors found encouraging evidence that the compounds could produce functional improvements in models of Huntington's disease across three species. First, patient-derived neurons treated with each of the compounds showed significantly less shrinkage, degeneration of neuronal projections and cell death than was seen in untreated neurons. Second, flies that model Huntington's disease and were treated with the compounds recovered climbing ability and survived longer than did untreated counterparts. Third, treated mice that model Huntington's disease showed improvements in three motor tests, compared with untreated mice. That said, preclinical trials in mice will be necessary to ascertain that the benefit is sustained and robust over the course of long-term therapy.

Finally, Li et al. analysed mutant ataxin-3, a protein that is involved in a neurodegenerative disorder called spino-cerebellar ataxia type 3. The researchers found that the compounds targeted the long polyglutamine tract of mutant ataxin-3 and lowered protein levels. We already know that small reductions in the levels of mutant ataxin-1, ataxin-2 and ataxin-3 can reduce the severity of spino-cerebellar ataxia types 1, 2 and 3, respectively, in mouse models of 1. Thus, this therapeutic strategy might be useful not only for Huntington's disease, but also for other diseases involving expanded polyglutamine tracts.

Moving forwards, there are three major research paths to pursue. The first involves establishing the mechanism by which Li and colleagues' compounds recognize proteins with expanded polyglutamine tracts but spare normal proteins. Perhaps the compounds recognize a particular structural conformation that arises only after the polyglutamine tract exceeds a specific length. The second involves testing the compounds in other models of polyglutamine disorders and assessing their effects.

The third path involves conducting similar

small-molecule screens for compounds that can clear polyglutamine proteins using other types of protein-clearance machinery. For instance, small molecules dubbed proteolysistargeting chimaeras (PROTACs) link a ubiquitin ligase enzyme to a protein of interest. The enzyme tags the protein with ubiquitin groups, leading to the protein's degradation by a cellular machine called the proteasome<sup>12</sup>. PROTACs have yet to be applied to a polyglutamine-expanded protein. But given that some of these proteins are degraded by the proteasome, the strategy could well prove viable — as long as it targets only the abnormally long polyglutamine tract.

**Huda Y. Zoghbi** is in the Departments of Molecular and Human Genetics, Pediatrics, Neurology, and Neuroscience, Baylor College of Medicine, Houston, Texas 77030, USA, and at the Jan and Dan Duncan Neurological Research Institute at Texas Children's Hospital, Houston, and is an Investigator with the Howard Hughes Medical Institute. e-mail: hzoghbi@bcm.edu

- 1. Li, Z. et al. Nature **575**, 203–209 (2019).
- Ravikumar, B. & Rubinsztein, D. C. Mol. Aspects Med. 27, 520–527 (2006).
- Lackey, K. et al. Bioorgan. Med. Chem. Lett. 10, 223–226 (2000).
- Reddy, K., D'Orazio, A., Nadler, E. & Jain, V. K. Clin. Genitourin. Cancer 4, 156–159 (2005).
- 5. Liu, J. P. & Zeitlin, S. O. J. Huntington's Dis. **6**, 1–17 (2017).
- Kordasiewicz, H. B. et al. Neuron 74, 1031–1044 (2012).
   Spronck, E. A. et al. Mol. Ther. Methods Clin. Dev. 13.
- 7. Spronck, E. A. et al. Mot. Ther. Methods Clin. Dev. 1; 334–343 (2019).
- 8. Gennarino, V. A. et al. Cell **160**, 1087–1098 (2015).
- Friedrich, J. et al. JCI Insight 3, e123193 (2018).
- 10. Scoles, D. R. & Pulst, S. M. RNA Biol. 15, 707-714 (2018).
- 11. McLoughlin, H. S. et al. Ann. Neurol. 84, 64-77 (2018).
- Sakamoto, K. M. et al. Proc. Natl Acad. Sci. USA 98, 8554–8559 (2001).

This article was published online on 30 October 2019.

#### **Engineering**

# Soft microbots controlled by nanomagnets

#### **Xuanhe Zhao & Yoonho Kim**

Arrays of nanoscale magnets have been constructed to form the magnetized panels of microscopic robots — thus allowing magnetic fields to be used to control the robots' shape and movement. See p.164

In science-fiction films, robots are often depicted as human-sized or larger machines made of rigid materials. However, robots made of soft materials or with flexible structures. and that can be much smaller than the human body, have attracted great interest in the past few years because they have the potential to interact with humans more safely than can rigid machines. Indeed, sufficiently small soft robots could even be used for biomedical applications in the human body. Various options are available to power these robots, but magnetic fields offer a safe and effective means of wireless operation in confined spaces in the body. On page 164, Cui et al. report a key step towards the fabrication of micrometre-scale robots that, in a programmable manner, can quickly morph into different shapes in applied magnetic fields.

The ability of minerals known as lodestones to align with Earth's magnetic field was first reported in the ancient Chinese manuscripts *Gui Gu Zi* and *Han Fei Zi*, and was later used in early magnetic compasses<sup>2</sup>. A similar principle has been used in the past few years in magnetic soft robots<sup>3–10</sup>, in which magnets of varying sizes (nanometres to millimetres)

are integrated into flexible structures or soft materials. The tendency of the magnets to orient in externally applied magnetic fields provides a way of quickly moving or changing the shape of these untethered robots remotely. This actuation mechanism allows much flexibility in the design of the robots' structures, magnetization patterns and strengths, and in when and where magnetic fields are applied to control the robots. In addition, because the forces and torques exerted on magnets by external magnetic fields can be accurately calculated, models have been developed to quantitatively describe the actuation of specific robot designs<sup>11</sup>.

Magnetic soft robots have been developed for various uses, especially in biomedical applications in which they interact closely with the human body. For example, self-folding 'origami' robots have been reported that can crawl through the gut, patch wounds and dislodge swallowed objects<sup>4</sup>; and capsule-shaped robots have been made that roll along the inner surface of the stomach and can perform biopsies and deliver medicine<sup>3</sup>. Magnetically steerable robotic catheters have also been developed, which can perform minimally

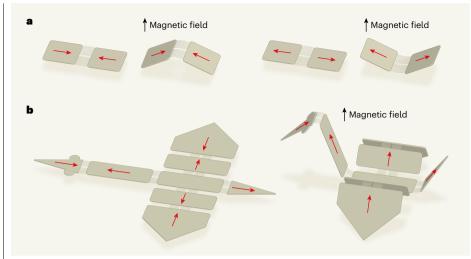


Figure 1 | Magnetic soft microbots morph on cue. a, Cui et al. have fabricated microscopic components consisting of magnetized panels connected by flexible hinges. When an external out-of-plane magnetic field is applied, the panels move in a direction that depends on the panels' direction of magnetization (red arrows) and on the direction of the applied field. For example, this two-panel system bends at the hinge. **b**, Robots assembled from panels that have different magnetization directions can thus be made to undergo complex movements when a sequence of magnetic fields is applied, such as this bird producing flapping movements.

invasive surgery on the heart or inspect lung airways<sup>5,7</sup>. And much thinner, thread-like robots have been made that could potentially navigate the brain's blood vessels to treat strokes or aneurysms<sup>10</sup>. These robots range in size from hundreds of micrometres to a few centimetres in diameter.

Further miniaturization of magnetic soft robots could enable new applications, such as performing operations in the smallest blood vessels and manipulating single cells, but the fabrication of such tiny machines poses a considerable challenge. Existing methods for the construction of small magnetic soft robots have included the direct assembly of magnetic components<sup>3-5,7</sup>, the magnetization of particle-loaded polymer sheets6, and the printing of soft composite materials that contain aligned magnetic particles<sup>9,10</sup>. Cui and colleagues now push the technological boundaries further, by using a technique called electron-beam lithography to make magnetically reconfigurable robots at scales of just a few micrometres. More specifically, this technique enables them to prepare arrays of nanoscale cobalt magnets in panels on a thin, flexible substrate of silicon nitride (Si<sub>3</sub>N<sub>4</sub>).

The authors' cobalt nanomagnets can retain their magnetism after exposure to an external magnetic field. This behaviour is called hysteresis, and results, in part, from the nanomagnets' shape. The authors could therefore tune the nanomagnets' magnetic properties and hysteretic behaviour so that thinner nanomagnets were harder to magnetize than thicker ones; in other words, stronger magnetic fields were required to magnetize thinner nanomagnets. This, in turn, meant that it was easier to re-magnetize thicker magnets - to 'over-write' the strength and direction of their magnetization – using relatively weak

Cui and colleagues could therefore selectively tune the magnetization of the nanomagnets so that an actuating magnetic field (much weaker than the fields that initially magnetized them) caused different panels to fold in different ways. The resulting multi-panelled components were thus 'programmed' to morph into specific configurations in an actuating magnetic field (Fig. 1). These components could, in turn, be assembled to produce complex shapes, such as letters, and even to make a microscopic 'bird' that produces motions such as turning, flapping and slipping across a surface.

### "The authors made a microscopic 'bird' that flaps, turns and slips across a surface."

Much work must still be done to achieve the full potential of magnetic soft robots for biomedical applications across various length scales. They must be designed using quantitative models to optimize their performance for specific tasks in relatively weak magnetic fields - that is, to work out which reconfigurations are needed, the sizes of the forces that the robot must exert on its environment, and the speeds at which reconfigurations should occur and with which the forces should be applied. Advanced fabrication platforms, such as the one used by Cui et al., will be crucial for implementing future designs.

Methods for the real-time imaging and

localization of robots deep in the human body are also needed, particularly in tight spaces. and must not interfere with the magnetic-actuation mechanisms. Artificial intelligence might be further developed to assist image analysis and robot control. Lastly, methods are needed for the safe retrieval or degradation of robots once they have performed their tasks. Degradation without toxicity or other adverse effects is particularly desirable.

Magnetic soft robots are also being extensively studied for applications beyond biomedicine8, such as in flexible electronics, reconfigurable surfaces and active metamaterials (engineered materials consisting of subunits that take in energy locally, and then translate it into movement that can produce large-scale dynamic motion). A parallel set of platforms for the design, fabrication, imaging and control of magnetic soft robots across various length scales are therefore under development. That work, together with developments such as those of Cui and colleagues, is laying the foundation for this nascent field.

Xuanhe Zhao and Yoonho Kim are in the Department of Mechanical Engineering, Massachusetts Institute of Technology. Cambridge, Massachusetts 02139, USA. e-mail: zhaox@mit.edu

- Cui, J. et al. Nature 575, 164-168 (2019).
- du Trémolet de Lacheisserie, É, in Magnetism: Fundamentals (eds du Trémolet de Lacheisserie, É., Gignoux, D. & Schlenker, M.) 3-6 (Springer, 2005).
- Yim, S. & Sitti, M. IEEE Trans. Robot. 28, 183-194 (2012).
- Miyashita, S. et al. 2016 IEEE Int. Conf. Robot. Automat. 909-916 (IEEE, 2016).
- Edelmann, J., Petruska, A. J. & Nelson, B. J. J. Med. Robot. Res. 3, 1850002 (2018)
- Hu, W., Lum, G. Z., Mastrangeli, M. & Sitti, M. Nature 554, 81-85 (2018)
- Jeon, S. et al. Soft Robot. 6, 54-68 (2019).
- Kim, Y., Yuk, H., Zhao, R., Chester, S. A. & Zhao, X. Nature **558**, 274-279 (2018).
- Xu, T., Zhang, J., Salehizadeh, M., Onaizah, O. & Diller, E. Sci. Robot. 4, eaav4494 (2019).
- 10. Kim, Y., Parada, G. A., Liu, S. & Zhao, X. Sci. Robot. 4, eaax7329 (2019).
- 11. Zhao, R., Kim, Y., Chester, S. A., Sharma, P. & Zhao, X. J. Mech. Phys. Solids 124, 244-263 (2019).

#### **Cancer genetics**

# **Genomes captured during** tumour spread

#### Jillian F. Wise & Michael S. Lawrence

A better understanding of the genetic changes that enable cancers to spread is crucial. A comprehensive study of whole-genome sequences from metastatic cancer will help researchers to achieve this goal. See p.210

The major cause of cancer-related deaths is the spread of cancer cells from their primary site to other parts of the body1. This spreading process, known as metastasis, typically involves cellular stressors and environmental shocks that induce dramatic changes in cancer cells. One such change is a fierce resistance to current therapies, which means that new ways to combat metastatic disease are urgently needed. On page 210, Priestley et al.2 use whole-genome sequencing (WGS) to illuminate the genomic changes that underpin metastasis in 22 types of solid tumour. Although previous studies<sup>3,4</sup> have unearthed some hints of such changes, this is perhaps the first pan-cancer metastasis study of its size to exploit the power of WGS.

Priestley et al. characterized 2,520 samples of metastatic tumours from people with a sample of non-cancerous blood cells from the same person. Using WGS, the authors produced a rich catalogue of the genetic mutations found in each metastasis. This catalogue complements existing inventories from both metastasis-sequencing studies and genomic databases of primary tumours, and offers several interesting insights. For example, the authors reveal frequent mutations in the gene MLK4; this is consistent with a previous study that connected an increased number of copies of MLK4 with metastasis5.

Most of the authors' findings confirm previous work on metastatic cancers<sup>3,4</sup>. For instance, other studies did not find recurrent cancer-causing mutations that were specific to metastatic tumours (that is, absent in the primary tumour) and that thus might

cancer (Fig. 1). In each case, they also analysed

Metastatic Primary tumou Oncogene mutation mutation Deletion Whole-genome doubling

Figure 1 | Characteristics common across metastatic cancers. Cells in a primary tumour typically harbour cancer-causing mutations (oncogenes). As the cancer evolves, it acquires further mutations that enable it to spread to other sites in the body through the blood – a process called metastasis. Priestley et al.<sup>2</sup> sequenced the entire genomes of 2,520 metastatic tumours, across 22 cancer types. They find frequent mutations in the gene MLK4. They also report widespread structural variations, such as whole-genome doubling (which they find to be especially common) and deletions of large chromosomal regions.

have triggered metastasis. This has led to speculation that, at least in solid tumours. metastasis-specific mutations are not the major cause of cancer spread<sup>1</sup>. Priestley et al. also found limited evidence of such mutations.

The researchers analysed not only single-nucleotide (point) mutations, but also large structural variations, including the deletion of DNA sequences and translocations of DNA from one chromosomal region to another. Structural variations are difficult to detect using sequencing techniques that cover small portions of the genome – sequencing of only protein-coding regions, for instance, or of even smaller targeted sequences. These techniques are used more frequently than WGS in clinical studies because of their affordability. Documentation of large structural variants is therefore a valuable feature of Priestley and colleagues' WGS study.

In particular, the report reveals pervasive whole-genome doubling (WGD), in which the entire chromosome inventory is copied. Priestley et al. find WGD in up to 80% of cases in certain types of metastatic cancer, whereas the phenomenon has been reported in only about 30% of primary tumours6. Linked to chromosomal instability, WGD can confer multidrug resistance to chemotherapy. Furthermore, it might provide a buffer for cancer cells against the detrimental effects on fitness caused by genomic instability, such as damaging mutations and losses of chromosomal segments7.

Although Priestley and colleagues present a landmark study, future efforts could benefit from researchers also sequencing each person's primary tumour. Doing this would have allowed Priestley et al. to generate a detailed reconstruction of how each cancer's genome evolved along the route to metastasis. To compensate for this limitation, the authors leveraged a large WGS study of primary tumours (the International Cancer Genome Consortium's pan-cancer analysis of whole genomes<sup>8</sup>). The researchers compared point mutations and small insertions and deletions between the two studies. These analyses largely confirmed a previous report of high genomic concordance between primary and metastatic tumours9. However, the comparison also revealed that the ten most commonly mutated cancer-causing genes in primary tumours are even more frequently mutated in metastatic tumours. Furthermore, larger DNA aberrations such as structural variations and WGD are significantly more common in metastases in most cancer types. These findings indicate that a hallmark of metastatic progression is ongoing and accelerating genomic instability.

Another caveat concerning this study, acknowledged by the authors, involves the use of fine-needle biopsies as the major sample-collection method. These biopsies gather cells from only a tiny subregion of a metastatic site. The authors report that, on average, more than about 93% of mutations detected in a given sample were present in every cell of that sample. This is in stark contrast to previous studies10, which have reported much higher levels of variation. The extreme homogeneity observed by Priestley et al. could, in principle, reflect the fact that only a few founding cancer cells colonized each metastasis, but might instead reflect the limited regional sampling achieved by the fine-needle biopsy method.

Future clinical studies of metastasis are likely to consider liquid biopsies as an alternative collection method. Liquid biopsies involve collecting samples of a person's blood and applying specialized laboratory techniques to isolate cancer-derived components, such as circulating tumour cells, circulating tumour DNA and released subcellular vesicles. This approach is less invasive than fine-needle or surgical biopsies. It also offers other advantages, including the ability to collect cells simultaneously from all metastatic cancer sites in the body (instead of just one), and to repeat sampling at multiple times during treatment, thereby providing dynamic temporal information about a cancer and its response to therapy. Liquid biopsies also enable researchers to document metastatic evolution at the DNA, RNA and protein levels in parallel<sup>11,12</sup>.

Ultimately, the true value of any research comes from improvements to treatment. To maximize the potential for clinical impact, Priestley and colleagues' data set is open-access. The authors have already accumulated more than 80 collaborative requests to investigate topics ranging from the possible presence of viral genetic material in the samples to the relationship between the sequences and patient drug responses (go.nature. com/2ommmn2). The data set is also being used to investigate whether any mutational variants involved in driving metastasis lie in regulatory DNA regions, and to enable efforts to deduce the anatomical origin of metastatic cancers diagnosed without a known primary-tumour site. Indeed, it is already powering exploration of these questions. The publicly available repositories are also being used in a Drug Rediscovery protocol<sup>13</sup>, in which patients with metastases who have exhausted standard therapies are matched with promising off-label treatments (anticancer medicines that have not been specifically approved for use against the person's type of cancer) on the basis of results from WGS.

Obtaining metastatic biopsies is not without risks to the patient, such as bleeding and infection. This is partly why sample collection has been so limited until now. Those who donated samples to this study have provided researchers with a valuable gift. It is hoped that the database will, in turn, provide the new insights and therapeutic strategies that are so urgently

Jillian F. Wise and Michael S. Lawrence are at the Massachusetts General Hospital Cancer Center and Department of Pathology, Harvard Medical School, Charlestown, Massachusetts 02129. USA, and at the Broad Institute of Harvard and MIT, Cambridge, Massachusetts. J.F.W. is also in the Department of Cancer Immunology, Institute for Cancer Research, University of Oslo, Oslo, Norway. e-mail: mslawrence@mgh.harvard.edu

- Lambert, A. W., Pattabiraman, D. R. & Weinberg, R. A. Cell 168, 670-691 (2017).
- Priestley, P. et al. Nature 575, 210-216 (2019).
  - Robinson, D. R. et al. Nature 548, 297-303 (2017).
- 7ehir A et al Nature Med 23 703-713 (2017) Marusiak, A. A. et al. Oncogene 38, 2860-2875 (2019).
- Bielski, C. M. et al. Nature Genet. 50, 1189-1195 (2018).
- Dewhurst, S. M. et al. Cancer Discov. 4, 175-185 (2014)
- Campbell, P. J., Getz, G., Stuart, J. M., Korbel, J. O. & Stein, L. D. Preprint at https://doi.org/10.1101/162784 (2019).
- Reiter, J. G. et al. Science 361, 1033-1037 (2018)
- 10. Granahan, N. & Swanton, C. Cell 168, 613-628 (2017).
- 11. Yu, M. et al. Science 345, 216-220 (2014).
- 12. Medford, A. J. et al. NPJ Precis. Oncol. 3, 18 (2019).
- 13. van der Velden, D. L. et al. Nature 574, 127-131 (2019).

This article was published online on 23 October 2019.

#### **Experimental physics**

## Progress on the proton-radius puzzle

#### Jean-Philippe Karr & Dominique Marchand

Atomic physicists and nuclear physicists have each made a refined measurement of the radius of the proton. Both values agree with a hotly debated result obtained by spectroscopy of an exotic form of hydrogen called muonic hydrogen. See p.147

The proton, discovered 100 years ago<sup>1</sup>, is an essential building block of visible matter. The nucleus of a hydrogen atom consists of a single proton, making this atom a suitable platform for determining the proton's intrinsic properties. One such property is the proton charge radius, which corresponds to the spatial extent of the distribution of the proton's charge. In 2010, a highly accurate measurement of the proton radius was made using spectroscopy of muonic hydrogen – an exotic form of hydrogen in which the electron is replaced by a heavier version called a muon<sup>2</sup>. However, the value obtained was almost 4% smaller than the previously accepted one<sup>3</sup>. Bezginov et al.4, writing in Science, and Xiong et al.5, on page 147, report experiments that could represent a decisive step towards solving this proton-radius puzzle.

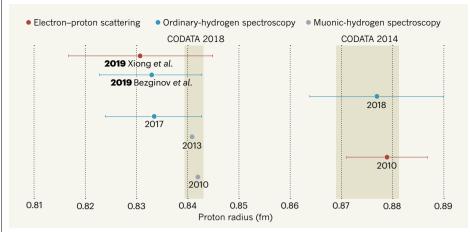
Atomic physicists determine the proton radius by measuring the energy difference between two electronic states of a hydrogen atom using spectroscopy. According to quantum mechanics, there is a non-zero probability that the electron will be found inside the proton if the electron is in a rotationless state (an S state). When inside, the electron is less strongly influenced by the proton's electric charge than it would otherwise be. This effect slightly weakens the binding of the electron and proton, and causes a tiny shift in the energy of the S state with respect to other states. The high precision achieved both by experiments and by the theory of quantum electrodynamics allows this energy shift and, in turn, the proton radius, to be extracted from measurements.

A muon is about 200 times heavier than an electron. As a result, there is a much higher probability that the muon in a muonic-hydrogen atom will be found inside the proton than would the electron in an ordinary hydrogen atom. Consequently, the associated energy shift is about 8 million (2003) times larger for muonic hydrogen than for regular hydrogen<sup>6</sup>. Muonic hydrogen is therefore a highly sensitive probe of the proton radius.

Bezginov and colleagues' work concerns the Lamb shift of ordinary hydrogen – the energy difference between the 2S and 2P excited states. This shift was investigated previously in muonic hydrogen<sup>2,7</sup>. To measure the Lamb shift, the authors developed an experimental method8 that derives from a technique known as Ramsey interferometry, which is used in atomic clocks.

This experimental method has many technical advantages over other approaches with regard to eliminating systematic uncertainties, filtering environmental noise, and simplicity in the shape of the spectral signal. A key feature of the set-up is the ability to measure a full spectrum in only a few hours. This allowed Bezginov et al. to carry out a meticulous study of systematic uncertainties and to extract a precise value for the proton radius: 0.833 ± 0.010 femtometres  $(1 \, \text{fm is} \, 10^{-15} \, \text{metres}).$ 

#### **News & views**



**Figure 1** | **Values for the proton radius.** A key property of the proton is its charge radius – the spatial extent of its charge distribution. This quantity is expressed in femtometres ( $1 \, \text{fm} \, \text{is} \, 10^{-15} \, \text{metres}$ ). The data points are values for the proton radius obtained over the past decade, including the latest results, from Bezginov et al.<sup>4</sup> and Xiong et al.<sup>5</sup>, with uncertainties indicated by the error bars. The data were obtained using three different measurement techniques: electron–proton scattering <sup>5,10</sup>, spectroscopy of ordinary hydrogen <sup>4,9,13</sup> and spectroscopy of an exotic type of hydrogen called muonic hydrogen<sup>2,7</sup>. The error bars for the two data points associated with muonic-hydrogen spectroscopy are too small to be depicted in this figure. The bands denote the values adopted by the Committee on Data for Science and Technology (CODATA) in 2014 (ref. 11) and in 2018 (see go.nature.com/2bwkrqz).

Nuclear physicists measure the proton radius using the 'elastic' scattering of electrons from protons. In this interaction, the incident electron transfers energy to the targeted proton through the exchange of a virtual (transient) photon. In a similar way to microscopy, short-wavelength photons (which transfer a lot of energy) reveal details at small scales. To determine the full extent of the proton's charge distribution, one should, in principle, use photons of infinite wavelength (that transfer zero energy), but no scattering at all would occur in this situation. Experiments therefore aim to achieve the lowest-possible energy transfer and then to extrapolate down to zero. This extrapolation, which relies on a parameterization of experimental data, is one of the main challenges in precisely determining the proton radius.

Xiong and colleagues implemented several key improvements over previous studies in their experiment, the Proton Radius experiment at Jefferson Laboratory in Virginia. Crucially, this investigation explores extremely low energy transfers (ten times closer to zero than previous data) while also probing larger energy transfers, to ensure consistency with existing data. The scattered electrons were detected through their energy loss in a detector called an electromagnetic calorimeter. This set-up avoided the need to use a magnetic spectrometer, the multiple settings of which induce systematic errors.

Furthermore, rather than making absolute measurements, Xiong *et al.* advantageously relied on relative measurements. Specifically, they determined the ratio between the number of events corresponding to elastic electron-proton scattering and the number related to

Møller scattering – a well-understood and calculable quantum-electrodynamics process in which electrons are scattered from atomic electrons. This strategy led to the cancellation of many systematic effects that are associated with absolute measurements.

In addition, the protons were in a hydrogen gas that was kept inside a chamber that did not have entrance and exit windows as used in previous similar experiments.

# "These independent measurements tip the scales in favour of a small proton radius."

This arrangement avoided background noise that would have been produced by the interaction of particles with window materials. Overall, Xiong and colleagues' chosen set-up, careful systematic-uncertainty checks at each step and exhaustive study of several parameterizations to extrapolate the data to zero energy transfer lend support to their value for the proton radius: 0.831 ± 0.014 fm.

The independent measurements of the proton radius made by Bezginov *et al.* and Xiong *et al.* are precise and consistent (Fig. 1). They tip the scales in favour of a small proton radius, in agreement with the highly accurate results from muonic-hydrogen experiments<sup>2,7</sup>.

But to conclusively solve the proton-radius puzzle, one still needs to understand why there are discrepancies between the latest results and the data from previous hydrogen-spectroscopy<sup>9</sup> and electron-proton scattering<sup>10</sup> experiments. For instance, the

value of the proton radius  $^{11}$  adopted by the Committee on Data for Science and Technology in 2014 was  $0.8751\pm0.0061$  fm. Because no convincing explanation for these discrepancies has been proposed, worldwide efforts must be pursued to validate the latest results and to critically assess the different measurement techniques.

Next-generation experiments will provide innovative approaches to this task. For example, the Muon Scattering Experiment<sup>12</sup> at the Paul Scherrer Institute in Switzerland is simultaneously investigating muon-proton and electron-proton scattering. This experiment is testing for possible differences in the behaviour of electrons and muons – an observation that would imply the existence of physics beyond that of the standard model of particle physics. On the spectroscopy side, high-precision measurements will be extended to other nuclei such as helium, and to molecules. It is highly probable that the harvest of results from future experiments will not only definitely solve, but might also explain, the proton-radius puzzle.

Jean-Philippe Karr is at the Laboratoire Kastler Brossel, Sorbonne Université, CNRS, ENS-Université PSL, Collège de France, 75005 Paris, France, and the Université d'Évry-Val d'Essonne, Université Paris-Saclay, France.

Dominique Marchand is at the Institut de Physique Nucléaire d'Orsay, CNRS-IN2P3, Université Paris-Saclay, 91406 Orsay, France.
e-mails: jean-philippe.karr@lkb.upmc.fr; dominique.marchand@in2p3.fr

- Rutherford, E. Phil. Mag. Ser. 6 37, 581–587 (1919).
- 2. Pohl, R. et al. Nature **466**, 213–216 (2010).
- Bernauer, J. C. & Pohl, R. Sci. Am. 310(2), 32–39 (2014).
- 4. Bezginov, N. et al. Science 365, 1007–1012 (2019).
- 5. Xiong, W. et al. Nature **575**, 147–150 (2019).
- Pohl, R., Gilman, R., Miller, G. A. & Pachucki, K. Annu. Rev. Nucl. Part. Sci. 63, 175–204 (2013).
- 7. Antognini, A. et al. Science 339, 417-420 (2013).
- Vutha, A. C. & Hessels, E. A. Phys. Rev. A 92, 052504 (2015).
- 9. Fleurbaey, H. et al. Phys. Rev. Lett. 120, 183001 (2018).
- 10. Bernauer, J. C. et al. Phys. Rev. Lett. 105, 242001 (2010).
- Mohr, P. J., Newell, D. B. & Taylor, B. N. Rev. Mod. Phys. 88, 035009 (2016).
- Gilman, R. et al. Preprint at https://arxiv.org/ abs/1709.09753 (2017).
- 13. Beyer, A. et al. Science 358, 79-85 (2017).

### nature

# insight

# Science and society

Editor, Nature

Magdalena Skipper

#### **Insights Editor**

Ursula Weiss

Science and society Insight Editor Anna Armstrong

#### Subeditors

Nicola Bailey, Sarah Farley, Dinah Loon, Mark McGranaghan, David Ribé, Francisca Schultz

#### Art Editors

Nik Spencer, Jennifer Burns, Denis Mallet, Claire Welsh

#### Production

lan Pope

#### Marketing

Steven Hurst

#### **Editorial Assistant**

Lyaina Vitalis

s illustrated by the 17 UN Sustainable Development Goals (https://www.un.org/sustainabledevelopment/), the challenges facing us are multifaceted and broad. The solutions will not lie in any one sector-all of society must engage. But science undoubtedly has a part to play in tackling pieces of the overarching puzzle of how to live in an equitable and healthy way that accords with the constraints of our planet and leaves no one behind.

In this collection of Reviews and Perspectives drawn together to celebrate 150 years of Nature, we take a look at a snapshot of areas in which science might help ease the transition to a healthy, sustainable and inclusive future.

In no way exhaustive, topics covered include the reuse and recycling of batteries from electric vehicles, the design and manufacture of more sustainable structural alloys, the capture and utilization of carbon in products, the resilience of harvestable ecosystems, the engineering of climate-resilient crops, progress and challenges in global vaccination, and epidemic prevention and response. Emphasised throughout is the need for a whole-of-society response if effective solutions are to emerge.

Science will serve us best if it accords with the principles of inclusivity, transparency and openness. In this vein, we end this Insight with a piece that shows how attending to one quite glaring blind spot in the scientific method-the overlooking of sex and gender-can open up new lines of enquiry and render research findings applicable to all.

There can be no division between science and society—even the purest and most abstracted lines of enquiry emanate from a social fabric and time, with its characteristic attitudes, limitations, blind spots and needs. Registering those restrictions and needs, and working to address them in a constructive and collaborative way, will help to keep the vision of a more humane and environmentally sound future, as laid out by the UN Sustainable Development Goals, alive.

Anna Armstrong, Rosamund Daw, Claire Hansell, Juliane Mossinger, Nonia Pariente, Sadaf Shadan, Clare Thomas, Ursula Weiss Senior Editors

#### Contents

64 Review Strategies for improving the sustainability of structural metals D Raabe et al

Recycling lithium-ion batteries from electric vehicles G Harper et al.

Perspective The technological and economic prospects for CO2 utilization and removal C Hepburn et al.

Perspective Anatomy and resilience of the global production ecosystem M Nyström et al.

109 Review Genetic strategies for improving crop yields J Bailey-Serres et al.

119 Review Immunization: vital progress, unfinished agenda P Piot et al.

130 Review A new twenty-first century science for effective epidemic response J Bedford et al.

137 Perspective Sex and gender analysis improves science and engineering C Tannenbaum et al.



Cover image Nik Spencer

The Campus 4 Crinan Street London N1 9XW. UK +44 (0) 20 7833 4000 nature@nature.com

# Strategies for improving the sustainability of structural metals

https://doi.org/10.1038/s41586-019-1702-5

Dierk Raabe1\*, C. Cem Tasan2\* & Elsa A. Olivetti2\*

Received: 24 April 2019

Accepted: 25 September 2019

Published online: 6 November 2019

Metallic materials have enabled technological progress over thousands of years. The accelerated demand for structural (that is, load-bearing) alloys in key sectors such as energy, construction, safety and transportation is resulting in predicted production growth rates of up to 200 per cent until 2050. Yet most of these materials require a lot of energy when extracted and manufactured and these processes emit large amounts of greenhouse gases and pollution. Here we review methods of improving the direct sustainability of structural metals, in areas including reduced-carbon-dioxide primary production, recycling, scrap-compatible alloy design, contaminant tolerance of alloys and improved alloy longevity. We discuss the effectiveness and technological readiness of individual measures and also show how novel structural materials enable improved energy efficiency through their reduced mass, higher thermal stability and better mechanical properties than currently available alloys.



Structural metallic materials have a historic and enduring importance in our society. They have paved the path of human civilization with loadbearing applications that can be used under the harshest environmental conditions, from the Bronze Age onwards. Only metallic materials encompass such diverse features as strength, hardness, workability, damage tolerance, joinability, ductility and toughness, often combined with functional properties such as corrosion resistance, thermal and electric conductivity and magnetism. This versatility comes with a vast understanding of thermo-mechanical processing of metals (accrued over millennia of metals use), which in turn enable numerous production, manufacturing, design, repair and recycling pathways.

#### Benefit and environmental impact of metallic alloys

Metals have enabled multiple applications in the fields of energy conversion, transportation, construction, communication, health, safety and infrastructure. Examples over the millennia have been agricultural tools, manufacturing machinery, energy conversion engines and reinforcements in huge concrete-based infrastructures. Recent applications include structural alloys for weight reduction combined with high strength and toughness in the transportation sector<sup>1-4</sup>, efficient turbines operating at higher temperatures for power plants and air traffic<sup>5,6</sup>, components for safe nuclear and fusion power and disposal<sup>7</sup>, targeted endurance or corrosive dissolution of biomedical implants<sup>8</sup>, embrittlement-resistant infrastructures for hydrogen-based industries9

or reusable spacecraft<sup>10</sup>. Metallurgical alloys and products boost innovation and economic growth: the global market for metals is about 3,000 billion euros per year<sup>11,12</sup>.

The success of the structural metals industry also means that it has an undisputable role in addressing our environmental crisis. The availability of metals (most of the elements used in structural alloys are among the most abundant), efficient mass producibility, low price and amenability to large-scale industrial production (from extraction to the metal alloy) and manufacturing (downstream operations after solidification) have become a substantial environmental burden: worldwide production of metals leads to a total energy consumption of about 53 examoules (10<sup>18</sup> I) (8% of the global energy used) and almost 30% of industrial CO<sub>2</sub>-equivalent emissions (4.4 gigatons of carbon dioxide equivalent, Gt CO2eq) when counting only steels and aluminium alloys (the largest fraction of metal use by volume)<sup>13</sup>; see Table 1. Although the production volumes of nickel and titanium are much smaller, they have an eminent role in aerospace and biomedical materials and nickel is primarily used as an alloying element in stainless steel (accounting for two-thirds of nickel's uses). The worldwide annual production in terms of mass, energy and CO<sub>2</sub> is presented in Table 1, with metal lost in manufacturing for these four key structural metals (where nickel use in stainless steel is the focus).

Mining and production of these materials have a huge impact in terms of resource use, emissions and waste generation, and this impact continues to grow, because of trends around urbanization, electrification and digitization (in 1950 less than 30% of the population lived in cities but this number is projected to exceed 60% by 2025). In addition, there are substantial byproducts of both industries that cause considerable environmental damage when not managed properly in perpetuity (losses throughout the supply chain are shown in Fig. 1 along with the quantities of material recovered as scrap). The two largest material groups (steel and aluminium) alone create huge mining and extraction byproducts, namely, 2,400 million tons (Mt) per year of tailings

Max-Planck-Institut für Eisenforschung, Düsseldorf, Germany. 2Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. \*e-mail: d.raabe@mpie.de: tasan@mit.edu: elsao@mit.edu

Table 1 | Overview of the energy and environmental impacts of key structural metals

	Worldwide annual production (Mt yr <sup>-1</sup> )	Energy (EJ yr <sup>-1</sup> )	CO <sub>2</sub> (CO <sub>2</sub> eq yr <sup>-1</sup> )	Material scrapped in manufacturing
Steel	1,700 (of which 45% is based on scrap input)	40	3.7 Gt	25%
Al	94 (of which 30% is based on scrap input)	13	0.7 Gt	40%
Ni (stainless steels/superalloys)	2.1 (of which 25% is based on scrap input)	0.25	26 Mt	20%
Ti	0.2 (limited post-consumer scrap)	0.07	6.7 Mt	60%

and 220 Mt per year of slags for steels and 160 Mt per year of bauxite residue for the case of aluminium.

Accidents, such as the iron-ore mining dam collapse in the mineralrich state of Minas Gerais, Brazil, in 2019 or the Aika, Hungary, spill in 2010 where 100,000 cubic metres of red mud breached a dam, show that these byproducts provide a constant threat and risk associated with extraction of the precursors to structural alloys. These energyconsumption challenges and detrimental environmental impacts are the biggest obstacle for further use of structural metals (Table 1 and Fig. 1).

To outline the critical opportunities towards more sustainable structural metals, this Review describes several approaches and measures. We discuss direct sustainability effects for different steps along the value chain including CO<sub>2</sub>-reduced primary production, secondary production through recycling and more efficient manufacturing (see 'Direct sustainability measures' and 'From geo-mining to urban mining' sections). In this context, we also discuss opportunities to make alloy design more recycling-oriented upfront (see 'Sustainable alloy design and recycling-friendly materials' section). Another strategy focuses on improved alloy longevity through corrosion protection, damage tolerance and repairability for longer product use (see 'Longevity by corrosion protection, lifetime extension and re-use' section). Finally, we illustrate how structural metals also enable energy-efficient products and processes indirectly through improved energy conversion and weight reduction in transportation at present or higher safety and lower costs (see 'Higher energy efficiency through lightweighting and harsher operating conditions' section).

Frequently, addressing environmental impact involves tradeoffs and undesired consequences (such as vehicles not becoming more fuel-efficient despite materials innovation and the use of lightweight alloys because extra performance and luxury items are added). To ensure that we are improving sustainability outcomes, one must consider the environmental impact of these strategies quantitatively to avoid unintended consequences, wherein we improve one aspect at the detriment of another part of the material system or life cycle. Furthermore, one must consider the economic feasibility, the technology readiness of potential solutions, and the role of governmental legislation. Such legislation could mean that metal production will be limited by actions to limit greenhouse gas emissions. The scope of our discussion here primarily focuses on the technical aspects of proposed solutions, but we indicate the feasibility and viability of these opportunities, that is, we assess the effects that the different measures can have on enhancing the sustainability of structural alloys.

Figure 2 presents these critical opportunities along two axes, their scaled potential for impact versus their technology readiness. In Fig. 2 we qualitatively rank the impact of each of the strategies, in addition to how soon the impact might occur, based on the status of the technology (and societal willingness to adopt the approach). Each strategy is also differentiated by the metal industry where it may have the most impact (or not, if the potential impact is for all alloys). For the lower-volume alloys, containing mainly titanium and nickel, the qualitative impact potential is scaled by the size of the industry. For example, reduced scrap in manufacturing holds considerable potential for sustainability improvements within titanium-related value chains (green in Fig. 2) even though reductions in manufacturing scrap would have a higher

impact overall for steel, given its larger production volumes. Irrespective of the qualitative nature of Fig. 2 and the subjective placement of each strategy, we offer this as a framework within which to understand the relative potential of each option.

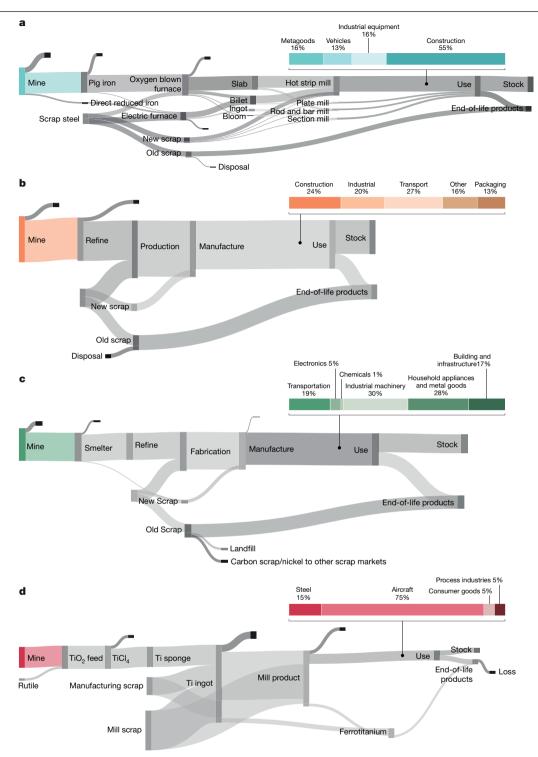
#### **Direct sustainability measures**

Improving the direct sustainability of structural alloys refers to reducing the environmental footprint of production and manufacturing. Ideally, moving towards more sustainable materials can be coupled with improvement in the material's performance and longevity. In this section we focus on the opportunities within production first, and then within manufacturing.

As shown in Fig. 1, large fractions of metal still flow into societal stock (the infrastructure and products that we use), so efforts must focus first on production, where there is the most potential for reduction of carbon dioxide emissions. This means that recycling alone cannot address production efficiency because the world's demand for metallic alloys exceeds the available amount of scrap by about one-third, at least up to 2050<sup>13–16</sup>. Improvements in production will vary by metal, except that all production processes would benefit now from use of non-carbon-intensive energy sources and better byproduct management (particularly wastes from steel and aluminium), as we suggest in Fig. 2. These strategies are more technology-ready than CO<sub>2</sub>-reduced approaches. Another important approach across production of all metals (also shown in Fig. 2) lies in harvesting the enormous waste heat from metal production, which could be used for electricity production. We focus primarily on steel in our discussion on CO<sub>2</sub>-reduced production, with only a brief mention of aluminium, given that steel is where most of the opportunity exists (Fig. 2).

For low- and medium-alloyed carbon steels, coke-making, blast furnace operation and steel-making account for the largest fraction of CO<sub>2</sub> emissions (2 billion tons) along the metal value chain, producing about 5.5% of the world's total CO<sub>2</sub> emissions from all fossil fuel burning<sup>17</sup>. A process with greatly reduced process emissions can be realized through electrolytic reduction of iron oxide in an alkaline solution at 110 °C and subsequent processing in electric arc furnaces, realizing a fully electrified synthesis route. This approach is a disruptive alternative to carbon reduction in the blast furnace because it uses only electrical rather than fossil fuel as the energy carrier (this electricity must then come from renewable sources)<sup>18</sup>. Up until now, wider use of this process has been impeded by the high costs and the aggressive conditions to which the electrodes are exposed. Technology readiness studies assume that electrolytic iron synthesis, which has not yet reached pilot-plant scale, is unlikely to enter the market before 2040<sup>19</sup>.

There are also steps that can be taken to render blast furnace and converter production more sustainable. A reduction in CO<sub>2</sub> emissions of up to 30% can be reached through (1) the addition of hydrogen to fossil reducing agents (such as coal) in the blast furnace, which also increases efficiency and production rate as a result of hydrogen's high percolation rate<sup>20,21</sup>; and (2) CO<sub>2</sub> capture and the downstream catalytic reduction of the waste CO<sub>2</sub> into alternative chemical products and/or energy carriers<sup>22,23</sup>. These techniques are available and are currently being studied at pilot-plant scale. Whereas CO<sub>2</sub> capture is ready to enter the market (depending on investment, carbon taxation and political decision-making to sanction this technique), the use of hydrogen in



 $\textbf{Fig. 1} | \textbf{Global material flows along the supply chain.} \, \texttt{Data from} \, 2017^{150} \, \texttt{are}$ shown for iron and steel<sup>151</sup> (a), aluminium<sup>150</sup> (b), nickel<sup>44</sup> (c) and titanium<sup>46</sup> (d). The width of the flows corresponds to mass at each stage including metal loss

(not to scale). The inset charts show dominant end uses for these materials. This excludes use of titanium in the pigment sector (at present the largest sector at around 90%) because our emphasis is on metals.

existing blast furnaces and downstream chemical conversion of CO<sub>2</sub> into value-added chemical products are less mature technologies, the first one owing to safety and infrastructure issues and the second one owing to the impurity of the exhaust gas mixture, which renders the required catalysis processes challenging.

An alternative to the blast furnace is the solid-state reduction of ores. In these direct reduction methods, porous iron-oxide fillings are reduced into pelletized aggregates with >95% Fe content without going through a liquid phase<sup>24</sup>. Traditionally, the reduction agents have been carbon carriers (such as methane) in this process. Nowadays, more complex gas mixtures can also be used, containing hydrogen, carbon monoxide and/or methane. Compared to traditional blast furnace ironmaking, CO<sub>2</sub> emissions can be reduced through direct reduction by up to 50% depending on the hydrogen content of the used gas mixture. Hydrogen then acts as a reductant along with the carbon monoxide.

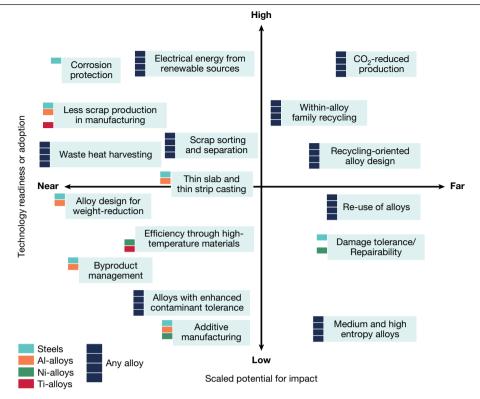


Fig. 2 | Impact and technology readiness of sustainability measures for structural alloys. We present an overview of several strategies and suggested potentials for impact on the sustainability of structural metals. Colours correspond to metals for which the opportunities are greatest (blue for carbon

steel, orange for aluminium, green for nickel alloys and stainless steels, red for titanium). Where no colours are shown, all metals provide similar opportunities for that strategy. We include the qualitative potential (weighted by the relative scale of each industry) for each strategy by metal.

Hydrogen-enriched direct reduction is already available at industry scale. The competitive market entry of fully hydrogen-based direct reduction techniques without any CO<sub>2</sub> emissions is currently being explored and expected to enter the market around 2030<sup>25</sup>. The primary reduction of aluminium is done via electrolysis so the main opportunities for improving its sustainability involve using renewable electricity sources (a strategy that would benefit all metals), improved efficiency of the electrolysis cells, and lower consumption rates of the currently used graphite or alternative-material electrodes<sup>26</sup>.

We next shift our attention downstream to manufacturing. In the downstream manufacturing following primary production, the overall yield losses occurring through liquid metal processing, forming and fabrication of aluminium and steel are 40% and 25% by mass, respec $tively ^{27,28}. \ These \ arise \ primarily \ from \ challenges \ involving \ the \ form \ of$ upstream products, the nature of upstream processing, the surface finish requirements, the supporting materials needed for shaping, and defects. Energy savings based on eliminating metal loss are estimated at around 5% and 15% for aluminium and steel, respectively. Alloy-specific high-quality material recovery already occurs inside casting and rolling plants where closed-loop procedures are established. Only a few producer-customer groups have established closed chains of returning alloy-specific scrap, so there are still substantial opportunities here that could be aided by data-driven approaches to process control and scrap sorting (see section 'From geo-mining to urban mining').

Near-net-shape manufacturing methods, where parts are cast or printed with a shape close to the final shape (thus requiring less machining) may provide an opportunity to reduce manufacturing material losses. Yet, additive manufacturing is not itself a sustainable synthesis approach owing to the high loss of powder, which can only be re-used a few times before it gets too oxidized (see Fig. 2 for the potential benefits of this strategy, most relevant for aluminium (Al), nickel (Ni) and titanium (Ti) alloys and tool steels)<sup>29</sup>. However, large-scale near-net-shape manufacturing methods have advanced. Examples are thin-slab and thin-strip casting of steels and aluminium (medium-term readiness and scaled impact; see Fig. 2). In conventional casters, steel slabs have a thickness of 150-300 mm. The usual target of a hot band thickness of 2-3 mm thus requires >99% mechanical working, resulting in high energy and investment costs. Industrial thin-slab technology can provide slabs of thickness 25–60 mm. For some steel grades, higher scrap fractions and higher contaminant content can be tolerated when casting thin slabs, owing to the high solidification rates (resulting in less segregation and fewer intermetallics forming). The next step is thinstrip steel casting<sup>30,31</sup>. In this approach, liquid steel is cast between two rolls and exposed to a hot reduction step to manufacture strips 2-3 mm thick that are directly coilable. The reduction in plastic work, energy and investment is enormous, yet production speed is slow and thin strips often do not reach the surface quality demanded by the market. Similar trends apply to strip production of aluminium alloys 32,33. Several belt and twin roll techniques were developed, which are capable of casting aluminium strips with thickness 1-15 mm. Specific challenges lie in the resulting centre-line segregation and the associated formation of coarse intermetallics, which are particularly caused by typical scrap contaminants such as iron (Fe) and silicon (Si). The manufacturing of high-quality and high-strength automotive grades, as a typical highend mass product, remains challenging.

#### From geo-mining to urban mining

One of the most efficient direct sustainability methods of reducing energy consumption and emissions lies in offsetting some primary extraction from geo-mined resources by urban mining of scraps and remelting them into new structural alloys. This can involve several specific strategies (outlined in Fig. 2) varying in readiness from scrap sorting and separation, to within-alloy family recycling (our preferred term for closed-loop recycling), alloy design for weight-reduction and

#### **Review**

recycling-oriented alloy design. All have considerable opportunity to achieve a sustainable impact provided the market is responsive enough to offset some primary extraction<sup>34</sup>. The ability to do this effectively varies by alloy family and the source of the material (that is, whether it is post-industrial or post-consumer). By some measures for steel. particularly for the European Union and the USA, scrap availability is beginning to meet steel demand, so it will become critical to avoid contamination in recovery. Composition-sensitive steel recycling is a strategy that will offer more value than scrap-compatible steel design. Recovering post-consumer scrap requires the most sophisticated management practices and has the most room for improvement <sup>14</sup> given that new material made from such scrap currently requires dilution with a large proportion of primary synthesized material in order to obtain the required composition. As the demand for products that can act as impurity sinks starts to decrease, we will have a smaller overall potential for recovery. Steel is typically recovered into the construction industry, a saturated market in some regions, and aluminium is recovered into cast products, so as the cast market becomes replete with scrap, the demand for cast products becomes a recovery limit<sup>35</sup>.

Advanced sorting moves the downcycling that is prevalent in metal recovery from the scrap yard to the plant floor. Automated probing and sorting methods have traditionally suffered from high costs resulting from low throughput and these costs are tightly constrained by the margins within recovery industries. Given the variety of type, shape, size, and form of scrap material, it has proved challenging to developing a wide array of broadly applicable recycling technologies. Such technology needs to cover identification, size reduction, separation, cleaning and material liberation. Separation steps include magnetic, sieving and air separation along with density separation, eddy-current, and spectroscopic techniques<sup>36</sup>. For these latter steps and even in scrapyard inspection, elemental analysers based on X-ray fluorescence and laser-induced breakdown spectroscopy have increasing potential. For instance, for certain alloys such as the more composition-sensitive aluminium-silicon-magnesium (Al-Si-Mg) alloy class, techniques such as aluminium alloy separation by laser-induced breakdown spectroscopy are promising ways to screen and sort for specific doping elements such as iron and copper (Cu). As alloys become more diverse, detection methods more sensitive, and throughput increases, these methods will become economically competitive<sup>37</sup>. When using such methods, the associated energy costs must be considered when comparing recycling to primary production<sup>38</sup>.

Until scrap sorting is perfected, the challenge for the use of secondary materials in a broader range of products is impurity tolerance, particularly as increased use of scrap may lead to compositional drift of alloy streams<sup>39</sup>. This compatibility and potential for drift is particularly important for aluminium alloys, where the thermodynamics of the remelting processes dictate the fate of associated alloying and tramp elements, possibly leading to the formation of undesired intermetallic phases. Only around 20% of end-of-life scrap is turned into wrought aluminium, even though wrought products account for two-thirds of all aluminium in use<sup>40</sup>. The reason for this discrepancy is that many Al-Si alloys, used for cast products, are particularly tolerant regarding high scrap usage whereas ductile-sheet-forming variants based on the Al-Mg and Al-Mg-Si systems are very sensitive to impurities.

Steel recycling is primarily done in an electric arc furnace, but also about 10-20% of ore-based steelmaking is from scrap (used for cooling in an oxygen converter). Steel production in an electric arc furnace has lower total energy consumption, enabling more flexible use of scrap materials  $^{41}$ , but there is reduced flexibility regarding which alloys can be made because the refining reactions inside an electric arc furnace are challenging. One of the key limitations in steel recycling is that during the separation process there is incomplete separation of copper-containing components, an important contaminant in this downstream process. The copper content in a shredding process can be upwards of 0.25-0.3% (although, when copper prices rise, there is more incentive to manually

handpick this material out of the stream). Tin can also cause downstream processing problems, particularly in combination with copper <sup>42</sup>.

As mentioned above, the material cycle for nickel is closely linked to that of stainless steel, and this manifests in its recovery too. Stainless steel can be either ferritic or austenitic. The recycling rate is much higher for austenitic stainless steel (nickel-containing) than for ferritic stainless steel. When recycled through a shredder plant, the ferritic stainless steel portion will be collected magnetically together with the ordinary carbon-steel scrap and will therefore be mixed into carbon steel. The stainless-steel scrap mixed into carbon-steel scrap was estimated to have reached 32% of postconsumer stainless-steel scrap flows in 2010<sup>43</sup>. Studies have estimated that 80% of postconsumer nickel scrap is recovered within the nickel cycle, whereas 20% becomes a constituent of carbon and copper scrap (and is not recovered as nickel<sup>44</sup>). Therefore, improving nickel recovery requires both avoiding its dissipation in carbon steels and the separation of low-nickel from high-nickel austenitic steels. Both measures are important for reducing nickel loss and avoiding nickel contamination of recycled carbon steel<sup>45</sup>.

Given the focus on titanium within the aerospace sector, there is considerable potential for recovery from scrap generated in the production process (Fig. 2; for example 100 tons of titanium alloy scrap is generated in making a frame for a 787 aeroplane). Because of the strict specifications within the industries in which titanium is used, the long-lived nature of those products (translating to a lack of economic incentive based on lack of post-use volume), and challenges around oxygen and iron impurities, essentially all titanium recycling is post-industrial \*6. Post-industrial material can be re-introduced into the remelting step of the primary route to make titanium ingots.

The challenge with improved titanium recycling even for industrial scrap is oxygen contamination, which can be decreased by remelting scrap with virgin titanium, but only up to a limit (particularly if demand grows). Novel processing technologies are focused primarily on oxygen removal technology as well as management of flexible scrap remelting. Commercialized processes are focused on electrolytic refining of sponge titanium in molten salt and calcium deoxidation. Otherwise, less well sorted or more contaminated (for example, iron-contaminated titanium generated in the smelting process) titanium would be used predominantly within ferro-titanium (Fig. 1; global demand 60,000 tons per year) or as an additive within other metal streams such as steel, nickel, copper or aluminium.

Another fundamental challenge associated with high scrap usage in production is that the properties (such as strength, toughness or corrosion resistance) may vary intolerably between two different furnace charges. Scrap-dependent heat treatment adjustment and the required blending with primary material could be predicted using through-process computational materials engineering simulations that must be based on robust phase diagrams and kinetic databases. The customers primarily require specific properties rather than compositions, that is, it may be possible to correct composition-dependent property variations through adjusted, flexible and batch-specific downstream heat treatments, an area where data-driven methods might become helpful.

#### Sustainable alloy design and recycling-friendly materials

Progress and opportunities in recycling have focused on achieving an optimal fit of the collected and sorted scraps to existing alloys. Here we approach the task from a different perspective, namely, how the design of metallic alloys could be changed to render them more recycling-oriented upfront. This term refers to the capability of an alloy to be made from the highest possible fractions of (low-grade) scraps and at the same time to be compatible with other alloys when serving as scrap. This means that the elusive goal in optimized materials recovery is not only to understand better the influence of impurity elements on properties, but also to build recyclability directly into the design of materials. Current structural alloys are not devised for end-of-life but rather for one-time use. Research into developing a science of less-pure

and thus recycling-friendly alloys covers many aspects: small concentration thermodynamics and kinetics: impurity trapping at lattice defects: compositional existence ranges of phases; size, dispersion and composition of harmful intermetallic precipitates; solute-driven cohesion and decohesion effects and associated property changes. Optimization methods coupled with metallurgical design can suggest alloys whose compositional constraints can be modified towards more scrap-tolerant ranges while preserving performance<sup>39</sup>.

This approach marks a shift in alloy design, which currently aims in part at realizing new properties through changes in chemical composition. However, scrap compatibility in secondary production can be better realized when avoiding compositionally over-designed alloys and instead using materials from only a limited composition spectrum, where property tuning is achieved through microstructure adjustment. The best examples are steels, which provide hundreds of material variants with different microstructures and properties, yet all based essentially on the iron-carbon-manganese-silicon (Fe-C-Mn-Si) system.

When taking a closer look at the quest for recycling-friendly alloys, the two approaches (composition tuning and microstructure tuning) are not so different: while the approach of using only limited chemical variation holds, repeated processing of scraps leads to accumulation of contaminants in secondary synthesis. This turns simple alloy systems such as Al-Mg-Si (used, for instance, in automotive sheet production) into a multi-component<sup>47</sup> material containing also iron, manganese, chromium, titanium, zinc and copper. Some of these impurities can lead to brittle phases. This means that recycling-oriented alloy design must study the low-dose corners of multi-component phase diagrams that may take into account up to twenty elements. A better understanding of multi-component thermodynamics and kinetics is thus an important pillar of the design of more impurity-tolerant alloys<sup>48</sup>.

A related, but more disruptive, scrap-friendly alloy design approach consists in the design of crossover alloys, which are sometimes also referred to as broadband alloys or uni-alloys 49. This approach aims to replace the variety of over-designed alloys by a smaller number of materials each covering a broader range of properties to serve mass markets. For aluminium, where 250 specialized alloys are stocked but only 65 are regularly used, such crossover alloys could combine features of heattreatable and non-heat-treatable wrought alloys at broad composition tolerance and with wide application ranges, establishing a universal alloy concept. A specific example is the improvement of the strain-hardening capacity, which is needed in sheet forming. This can be achieved both by a higher solute magnesium content, owing to its effect on dislocation motion, but also by tuning crystallographic texture, reducing grain and dislocation cell sizes and improving precipitate dispersion.

Similar aspects apply to Al-Mn alloys, used in recycling-intensive branches such as packaging, which for non-safety-critical products can tolerate a large variation in composition (the impurity concentration of some contaminants such as iron, silicon, zinc and magnesium can vary by factors of up to five among batches; Fig. 3).

### Longevity by corrosion protection, lifetime extension and re-

Enabling longer-lasting products would reduce resource extraction through lifetime extension and repair ability of products or by facilitating re-use (Fig. 2)13. We note, however, that increasing product lifetime will not reduce the demand arising from increasing population. Therefore, for developing regions where the population is growing this strategy is of limited value. The longevity of alloys is mainly limited by fatigue, creep, corrosion including hydrogen embrittlement, thermal ageing and irradiation.

Fatigue is an effect of permanent microplastic deformation when a material is exposed to cyclic loading. This often occurs together with thermal and/or corrosive attack owing to the presence of oxygen and hydrogen, causing a phenomenon known as stress corrosion cracking. A similar scenario of gradual material decay is caused by creep,

which is a phenomenon where thermal activation enables plastic flow and microstructure coarsening of parts exposed to high homologous temperatures.

Corrosion and stress-corrosion cracking are by far the most severe phenomena limiting the longevity and integrity of metal products. destroying about 3.4% of the global gross domestic product every year, a value translating to US\$2.5 trillion (ref. 50) (see Fig. 2). Hence, any progress in corrosion resistance has large effects on longevity, most relevant for carbon steel.

In this context, loss of material and system failure due to oxidation accounts for the vast majority of the economic impact of corrosion and is an essential factor in infrastructure costs worldwide. Oxidation of metallic structures proceeds mostly through galvanic corrosion, which occurs when adjacent microstructural regions or different metals with unlike electrochemical potentials are in conductive contact. The electrochemically more active region then acts as anode and corrodes faster than the cathodic reactant. Galvanic corrosion is the prevalent decay mechanism when metal structures are in contact with an electrolyte such as water with solute ions.

Hydrogen embrittlement is another type of corrosion and poses a serious impediment for carbon-free hydrogen-propelled technologies. Unlike other corrosion products such as oxides and hydroxides, hydrogen is hard to detect and several embrittling effects can occur such as hydrogen-enhanced plasticity, decohesion, superabundant vacancies, hydride formation or nanovoids. The interplay among them makes it difficult to identify a clear cause of failure. Also, hydrogen-related damage can occur suddenly, causing abrupt catastrophic failure of structures. Hydrogen embrittlement can occur in structural alloys 51,52, particularly in iron, aluminium, nickel and titanium alloys with strength levels above 650 MPa.

Alloy lifetime can also be reduced by thermal and radiation effects<sup>53,54</sup>, causing brittle phases, enhanced abundance and mobility of lattice defects or capillary-driven microstructure coarsening 55,56. The industrial relevance of this is huge. Many components in the energy industry, specifically in nuclear reactors, can suffer from these phenomena, making it a field where safety issues can sometimes override sustainability concerns.

Measures to reduce fatigue and creep damage in alloys use some of the intrinsic damage-resistance and crack closure mechanisms that metals offer<sup>57–59</sup>. Examples are crack closure induced by plastic deformation, chemical reactions such as oxidation or athermal phase transformation 60-62 which are caused by the stress increase before a crack tip. triggering stress-driven phase transformation. This is often non-volume conserving, thus creating compressive stresses that can close crack tips. Another approach is diffusion-driven pore filling during creep<sup>63</sup>. Since most corrosion phenomena are interface-dominated reactions involving mass transport (mostly of metal and oxygen ions), corrosion protection (particularly against oxidation) is among the most important and efficient means of enhancing product longevity.

Corrosion protection methods are as varied as the underlying reactions and decay phenomena. Methods for mitigating oxidation depend on the underlying electrochemical reactions and the nature of the resulting products (see steel in Fig. 2). Countermeasures may rely on shifting the thermodynamic direction of the oxidation reaction by providing a sacrificial anode, engineering alloy compositions to favour formation of a protective oxide or to disfavour formation of detrimental phases, or directly preventing oxygen from reaching the vulnerable material via protective coatings. (Technological schemes, such as impressed current cathodic protection, are also widely applied but are beyond the scope of this review.)

Steel protection via zinc coatings is the most frequent application of a sacrificial anode against atmospheric oxygen exposure and accounts for half of the global zinc production of 13.5 Mt per year. Hence, environmental considerations apply also to zinc production and recycling when improving the longevity of steels. Some metals, such as many aluminium, titanium, nickel and stainless-steel alloys, form dense,

### Review

adherent, self-healing oxidation products that resist corrosion intrinsically by preventing further oxygen intrusion. Even in these materials. interfacial segregation and second phase formation can contribute corrosive damage. Finally, engineering protective coatings is an art in itself, since an imperfect or permeable coating can actually enhance the corrosion it is meant to prevent.

Protecting materials against hydrogen is challenging: whereas some materials such as titanium undergo formation of brittle hydrides, others such as nickel and high-strength steels experience enhanced local plasticity and void formation. Measures that may reduce hydrogen embrittlement are the reduction of regions of high micromechanical contrast among phases and microstructure components (as hydrogen tends to enrich in highly stressed regions such as at interfaces), dense oxide surface layers which reduce hydrogen uptake, and trapping of hydrogen at semi-coherent internal interfaces and supposedly also at other defects<sup>64-66</sup>. In some cases, hydrogen trapping can be a cause of undesired local softening, such as through the stabilization of superabundant vacancies or the lowering of the activation barriers for the double-kink mechanism of dislocation motion<sup>67,68</sup>.

When damage becomes visible, repairs can be done by cladding, welding or grinding for components ranging from bridges<sup>69,70</sup> to turbine blades<sup>71,72</sup>. Repairs can be conducted even when damage is not visible, for example, through maintenance treatments. There has been considerable research on self-healing metals, in which the main goal is that the material should have autonomous crack closure mechanisms<sup>73,74</sup>. Yet ambient temperature sluggishness of transformation kinetics in metals<sup>75,76</sup> limits the success of a truly autonomous crack closure mechanism to only a few case studies<sup>77</sup>. Most other cases require external treatments. The repairability of metals can be increased by focusing on removing microstructural changes that lead to embrittlement, instead of focusing only on micro-crack closure. This would enable prolonged utilization of intrinsic damage resistance that is intrinsic, but consumable (that is, it is gradually used up) in alloys exposed to repeated mechanical loads, which arise from plasticity  $^{62,78}$  or transformation  $^{58,79}$  micro-mechanisms. Thermally induced embrittlement effects in duplex stainless steels, which is due to G-phase formation (a CrNi-silicide), can be removed by annealing, enabling the part's re-use<sup>80</sup>. Similarly, cut-edge damage in sheet metal can be reduced by specifically designed cutting treatments<sup>81</sup>. Furthermore, such microstructure resetting strategies could enable further re-manufacturing processes for sheet metal that would increase re-use82.

Approaches to improving the sustainability of structural metals are supported by progress in computational methods. Metallurgists can now make use of databases of experimental data surrounding structure-property relationships, loading-specific precipitation, coarsening, phase transformation and even complete-lifetime predictions<sup>48,83,84</sup>. Data-driven approaches can use machine learning meth $ods^{85\text{--}89}$  that can sometimes be computationally more tractable than simulation-based approaches that aim to avoid damage-susceptible microstructures (to reduce failure) or to make predictions for when to apply repair treatments and how alloy compositions can be rendered more compatible for recycling.

Enabling re-use provides considerable opportunities for steel and aluminium, given that many applications in building and transportation reach end-of-life not because they fail but because they are replaced for economic reasons. Barriers to re-use are typically not technical in nature but rather economic, such as lack of demand, traceability concerns and lack of supply chain infrastructure 91. These systemic barriers need to be addressed to realize re-use potential through government leadership, education and information sharing.

#### Energy efficiency by lightweighting and harsh operating conditions

Metallurgical improvements can increase the performance and energy efficiency of industrial systems, products and processes simultaneously,

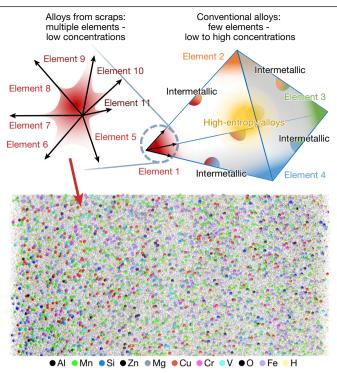


Fig. 3 | Multi-component allows for high scrap usage. Allows made from high scrap fractions can contain multiple contaminants, which means that they must be designed initially to be impurity-tolerant. This requires knowledge about low-concentration, multi-component phase diagram regions. The schematic shows a phase diagram with typical engineering alloys in the corners, where one element prevails, and intermetallics in the binary and ternary centres with fixed stoichiometries and high- and medium-entropy alloys as solid-solutions in the centres of the multi-component phase diagrams. The presence of scrap-related contaminants makes this a low-concentration, multi-component phase diagram. Below is an atom probe tomography dataset taken from a recycled Al-Mn base alloy used for packaging, revealing a high number of tolerable impurity elements, each below 0.1%. Such recycled Al-Mn alloys can also serve in infrastructure applications.

at reduced waste and greenhouse gas emissions. Here we tackle two pathways towards increasing 'in-use' energy efficiency: lightweighting (alloy design for weight reduction, Fig. 2) and increased operating temperature (efficiency through high-temperature materials, Fig. 2).

Worldwide, approximately 12% of steel products (about 121 Mt per year) and about 27% of aluminium products (12 Mt per year) are used in transportation, incentivizing efforts to reduce weight in automotive 92-95, aerospace<sup>96-98</sup> and railway<sup>99</sup> components. In the case of vehicle lightweighting, the potential is considerable given that 20% of our global energy and process CO<sub>2</sub> emissions originate from transportation and about 20% of that could be reduced through lightweighting 100. We note that, historically, vehicles have become heavier over time with improvements to comfort, performance and safety, even as our ability to lightweight the vehicle has improved. Lightweighting has sustainability benefits for other industries linked with transportation as well, such as the packaging (about 9 Mt per year of steel, about 6 Mt per year of aluminium) or construction industries (583 Mt per year of steel, 11 Mt per year of aluminium) (see Fig. 1). Efforts to reduce materials use through redesign, including alloy development and particularly microstructure tuning, could remove as much as 30% of steel and aluminium by mass from transportation uses.

In each of these industries, weight can be reduced (1) by using less metal, compensated by higher strength, (2) by using metal of lower density, or (3) by optimizing component design. (1) Using less metal means improved strength or elastic modulus properties must be achieved, while avoiding a decrease in toughness and ductility. There are several strengthening mechanisms that can be triggered in metals by optimizing

thermo-mechanical processing and/or composition, but many of them reduce ductility and toughness. Therefore, designing metals that exhibit simultaneous increases in these properties has been an essential goal of metallurgical research<sup>2,101-104</sup>. The most effective mechanisms for improving strength and ductility simultaneously have been by phase metastability. Transformation-induced plasticity (TRIP) and twinninginduced plasticity (TWIP) mechanisms have been extensively used for this purpose, in TRIP-assisted 93,105-107 or TWIP-assisted 108-110 steels and titanium alloys 111,112 through 'metastability' tuning of the stacking fault energy, achieved mainly by adjusting the carbon and manganese content and composition partitioning among phases. A more fundamental challenge is regarding the end product of strain-induced martensitic transformation: martensite. Although its transformation reaction has beneficial effects, the resulting fresh martensite and its bounding hetero-interfaces are often sites of damage nucleation 94. Similarly, the elastic modulus is an essential design consideration for vehicle mass reduction. Several high-stiffness metal matrix composites have been developed using stiff ceramic precipitates. Examples are Al-TiB2 and Fe-TiB<sub>2</sub>. Some of these materials provide improved stiffness of up to 10% with comparable formability. Upscaling depends on the capability to produce such alloys in larger quantities, for example, through in situ liquid metallurgy<sup>113</sup>.

(2) For the second approach to lightweighting, one goal is reducing the density of steels. When blending Fe with up to 25% manganese and up to 1.2% carbon the steel crystallizes into a face-centred-cubic structure, tolerating up to 8 atomic per cent aluminium in solid solution without formation of brittle phases. These materials are referred to as low-density steels and have a tensile strength up to 1.5 GPa at up to 60% elongation. When adding 20 atomic per cent Al to Fe-Mn-C alloys the mass density is reduced by as much as 15%, yet with reduced Young's modulus and precipitation of perovskitic carbides<sup>3</sup>.

Another avenue for weight reduction makes use of aluminiumbased<sup>4,114</sup> and magnesium-based<sup>1,115</sup> alloys. Current research is focused on achieving a better strength-ductility compromise for sheet applications, realizing ultrahigh-strength Al-Zn-Mg-Cu<sup>116</sup>, as well as weightreduced and stiffness-enhanced Al-Li alloys<sup>117</sup>. Several even lighter alloys are currently being developed, based on the Mg-Al-Zn115, Mg-Al-Ca<sup>118,119</sup> and Mg-rare-earth<sup>120</sup> systems, with a mass density as low as about 1.7 g cm<sup>-3</sup>, that is, about 80% reduced density compared to steels. Extreme weight reduction, but with insufficient corrosion resistance was realized in a Mg-Li alloy that approaches a mass density of only about 1 g cm<sup>-3</sup> (that is, that of water<sup>121</sup>).

(3) Design improvements can also lead to weight reductions. Here, our focus is not on topology optimization<sup>122</sup>, but on mesoscale optimization of the spatial distribution of microstructure features<sup>123</sup>. In cases where processing is feasible, grading of structure or composition can enhance properties<sup>124–127</sup>, enabling lightweighting. Materials designers have explored gradients of grain size 124,127-129, twin density 130,131 component or phase fraction<sup>132</sup>. These investigations demonstrated that the underlying physical micro-mechanisms of deformation can be influenced by such grading, leading to improvements in plastic $ity^{133,134}, strengthening^{135}, and \, damage \, resistance^{124,127,130,131,136}. \, However, \, and \, and$ processes that realize gradients are difficult to scale up (for example, accumulative roll bonding 137,138, thin film deposition 139,140). To this end, additive manufacturing methods offer potential for the synthesis of graded materials, and can be also employed to create functionally graded structures 141-146, although costs and production speed need to be improved.

Another field of interest is the overlap between additive manufacturing, alloy design, architectured materials, computational materials mechanics and bionics. Bionic design enables computer-generated topology-optimized lean geometries of parts at reduced mass and improved structural stiffness<sup>122,147</sup> (Fig. 4).

Another domain where structural alloys offer improved efficiency is the use of higher operational temperatures in energy conversion<sup>5,6</sup>.

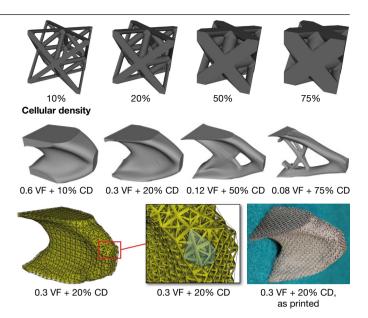


Fig. 4 | Synergies between structural optimization, additive manufacturing, architectured materials and bionics. Example of a three-dimensional cantilever beam design using these synergies. The upper row shows an architectured lattice structure with different cellular densities (CD) ranging from 10% to 75%. The middle row shows different combinations of volume fraction (VF) and CD for a part obtained by topology optimization. The four combinations all yield the same reduction in weight (94%) relative to the massive (un-architectured, not porous) cantilever beam with 100% VF and 100% CD. The image was compiled using results from ref. 147.

Such engines follow the Carnot cycle, and hence higher service temperatures yield better efficiency. As global electricity consumption amounts to 23 PW h (1015 W h) and electricity is the fastest-growing source of energy demand, most of it provided by turbines, higher conversion efficiency has huge potential for saving energy. This should stimulate research on nickel-based and cobalt-based superalloys, Tialuminides and Mo-Si-B alloys (Fig. 2).

#### Outlook on enhanced sustainability of structural metals

Structural metallic alloys have served as key enablers of human progress, wealth and wellbeing over millennia. Now, in the acceleration phase of the Anthropocene, their great advantages in terms of availability, mass producibility and low price have also become an environmental burden. Considering the huge quantities produced (1.7 billion tons of steel and 94 million tons of aluminium per year), the task of making structural alloys and their products sustainable is an enormous challenge. Metals-related sustainability solutions now require a holistic view of production and manufacturing, and a thorough metallurgical understanding of structure-property relations, product function and longevity, resource efficiency, pollution, market dynamics and societal impact<sup>12</sup>. These aspects are closely connected, with some of them being of thermodynamic nature (and therefore quantifiable), whereas others are harder to assess, such as customer response to green branding or market development<sup>148</sup>. We therefore recommend the use of marketinformed life-cycle assessment calculations for adequate risk and effect quantification before political and industrial decision-making, in order to reveal the true long-term efficiency gains of the various possible sustainability measures.

To improve the sustainability of metals production and use, several recommendations with both high leverage and high technologyreadiness follow from this overview. The greatest potential is attributed to the use of fossil-free and fossil-reduced energy sources in primary and secondary extraction and manufacturing as well as production methods that allow the dynamic use of green electricity. Also of great

#### Review

importance are improved corrosion resistance; materials-efficient manufacturing; improved product-to-product recycling; automated post-consumer scrap sorting; recycling-oriented alloy design and the development of multi-purpose crossover alloys.

A critical dimension in realizing many of the opportunities mentioned here that we have left untouched in our discussion is the importance of government action and economics. In the absence of regulations, the only driving force for emission reductions would be those strategies that can demonstrate economic benefit. The role of policy would be to leverage quantitative assessments to incentivize the most effective (in terms of sustainability) strategies<sup>19</sup>. Several current metal-industryrelated policies are inadequate, focusing on technology-specific deployment substitutes, or are not sufficiently transformative, therefore leaving potential for lock-in risk<sup>149</sup>. Measures that deserve immediate attention include pollution and emission controls across national borders, demand stimuli such as procurement mandates or recycling incentives that consider both sources and sinks for recovered metal, and supply measures that are technology-neutral but set benchmarks and standards for clean manufacturing while supporting scale-up needs, all coupled to supporting infrastructure and market analysis to enable the most economically viable strategies compatible with these sustainability goals<sup>152</sup>.

Systematically implementing the measures discussed will break with almost all traditions of our current industrial practice since the beginning of the first industrial revolution around 1800. Striving towards sustainability will become the next industrial revolution.

- Pollock, T. M. Weight loss with magnesium alloys. Science 328, 986–987 (2010).
   This milestone paper provides an overview of how to reach lightweight, energy-efficient, environmentally benign engineering systems using magnesium alloys
- Jiang, S. et al. Ultrastrong steel via minimal lattice misfit and high-density nanoprecipitation. Nature 544, 460–464 (2017).
- Raabe, D. et al. Alloy design, combinatorial synthesis, and microstructure-property
  relations for low-density Fe-Mn-Al-C austenitic steels. JOM 66, 1845–1856 (2014).
   This is an overview of aluminium-containing, high-manganese steels that are
  characterized by an excellent combination of strength and ductility, enabled by their
  high work hardening capacity.
- Hirsch, J. Aluminium in innovative light-weight car design. Mater. Trans. 52, 818–824 (2011).
- 5. Reed, R. C. The Superalloys (Cambridge Univ. Press, 2009).
- Pollock, T. M. & Tin, S. Nickel-based superalloys for advanced turbine engines: chemistry, microstructure and properties. J. Propuls. Power 22, 361–374 (2006).
- Bloom, E. E. The challenge of developing structural materials for fusion power systems. J. Nucl. Mater. 258–263, 7-17 (1998).
- Witte, F. et al. Degradable biomaterials based on magnesium corrosion. Curr. Opin. Solid State Mater. Sci. 12, 63–72 (2008).
- Luo, H., Li, Z. & Raabe, D. Hydrogen enhances strength and ductility of an equiatomic high-entropy alloy. Sci. Rep. 7, 9892 (2017).
- Dunn, B. D. Materials and Processes for Spacecraft and High Reliability Applications. (Springer, 2016).
- 11. Optimat Materials Landscaping Study. Report J2963/IUK (Optimat, 2018)
- World Economic Forum Mining & Metals in a Sustainable World 2050. Scoping Report (World Economic Forum, 2015).
- Allwood, J. M. & Cullen, J. M. Sustainable Materials: With Both Eyes Open (UIT Cambridge, 2012).
  - This seminal book is a very detailed and holistic treatment of all engineering and commercial issues related to materials sustainability with special attention paid to the required reduction of the industry's carbon emissions by 50% by 2050.
- Graedel, T. E. et al. What do we know about metal recycling rates? J. Ind. Ecol. 15, 355–366 (2011).
- Ferretti, I., Zanoni, S., Zavanella, L. & Diana, A. Greening the aluminium supply chain. Int. J. Prod. Econ. 108, 236–245 (2007).
- Mahfoud, M. & Emadi, D. Aluminum recycling—challenges and opportunities. Adv. Mater. Res. 83-86, 571-578 (2009).
- Strezov, V., Evans, A. & Evans, T. Defining sustainability indicators of iron and steel production. J. Clean. Prod. 51, 66–70 (2013).
- Allanore, A., Yin, L. & Sadoway, D. R. A new anode material for oxygen evolution in molten oxide electrolysis. Nature 497, 353–356 (2013).
- Fischedick, M., Marzinkowski, J., Winzer, P. & Weigel, M. Techno-economic evaluation of innovative steel production technologies. J. Clean. Prod. 84, 563–580 (2014)
- Sastri, M. V. C., Viswanath, R. P. & Viswanathan, B. Studies on the reduction of iron oxide with hydrogen. *Int. J. Hydrogen Energy* 7, 951–955 (1982).
- Chu, M., Nogami, H. & Yagi, J. Numerical analysis on injection of hydrogen bearing materials into blast furnace. ISIJ Int. 44, 801–808 (2004).
- Muramatsu, A., Sato, H., Akiyama, T. & Yagi, J. Methanol synthesis from blast furnace off gas. ISIJ Int. 33, 1144–1149 (1993).

- 23. Gielen, D.  $CO_2$  removal in the iron and steel industry. Energy Convers. Manage. **44**, 1027–1037 (2003).
- Battle, T., Srivastava, U., Kopfle, J., Hunter, R. & McClelland, J. The direct reduction of iron. In *Treatise on Process Metallurgy* Vol. 3 (ed. Seetharaman, S.) 89–176 (2014).
- Vogl, V., Åhman, M. & Nilsson, L. J. Assessment of hydrogen direct reduction for fossil-free steelmaking. J. Clean. Prod. 203, 736–745 (2018).
  - This book provides a detailed analysis of the feasibility and costs associated with the use of hydrogen as a reducing agent in the direct reduction of iron ores and the resulting opportunities to turn carbon-based iron production into a hydrogen- and electricity-based one.
- Gusberti, V., Severo, D. S., Welch, B. J. & Skyllas-Kazacos, M. Modeling the mass and energy balance of different aluminium smelting cell technologies. In *Light Metals 2012* (ed. Suarez, C. E.) 929–934 (2016).
- Olivetti, E. A. & Cullen, J. M. Toward a sustainable materials system. Science 360, 1396–1398 (2018).

### This paper discusses how materials extraction has shifted from Europe and North America to Asia and Africa.

- Allwood, J. M., Cullen, J. M. & Milford, R. L. Options for achieving a 50% cut in industrial carbon emissions by 2050. Environ. Sci. Technol. 44, 1888–1894 (2010).
- Azevedo, J. M. C. Cabrera Serrenho, A. & Allwood, J. M. Energy and material efficiency of steel powder metallurgy. *Powder Technol.* 328, 329–336 (2018).
- Raabe, D. Microstructure and crystallographic texture of strip-cast and hot-rolled austenitic stainless steel. Metall. Mater. Trans. A 26, 991–998 (1995).
- Raabe, D., Reher, F., Hölscher, M. & Lücke, K. Textures of strip cast Fe16%Cr. Scr. Metall. Mater. 29, 113–116 (1993).
- Haga, T., Nishiyama, T. & Suzuki, S. Strip casting of A5182 alloy using a melt drag twin-roll caster. J. Mater. Process. Technol. 133, 103–107 (2003).
- Monaghan, D. J., Henderson, M. B., Hunt, J. D. & Edmonds, D. V. Microstructural defects in high productivity twin-roll casting of aluminium. *Mater. Sci. Eng. A* 173, 251–254 (1993)
- Zink, T., Geyer, R. & Startz, R. A market-based framework for quantifying displaced production from recycling or reuse. J. Ind. Ecol. 20, 719–729 (2016).
- Rombach, G. Raw material supply by aluminium recycling-efficiency evaluation and longterm availability. Acta Mater. 61, 1012–1020 (2013).

## This work provides a detailed analysis and discussion of scrap cycles, scrap markets and associated secondary production opportunities for the case of aluminium alloys.

 Brooks, L., Gaustad, G., Gesing, A., Mortvedt, T. & Freire, F. Ferrous and non-ferrous recycling: challenges and potential technology solutions. Waste Manag. 85, 519–528 (2019)

### This review article provides detail regarding inbound inspection for metals recycling at a level of detail that enables action by industry and decision makers.

- Rombach, G. & Bauerschlag, N. LIBS based sorting—a solution for automotive scrap. In Light Metal 2019 (ed. Chesonis. C.) 1351–1357 (2019).
- Gaustad, G., Olivetti, E. A. Kirchain, R. Improving aluminum recycling: a survey of sorting and impurity removal technologies. Resour. Conserv. Recycling 58, 79–87 (2012)

### This paper carefully analyses the undesired accumulation of tramp elements during the recycling of aluminium alloys.

 Modaresi, R., Løvik, A. N. & Müller, D. B. Component- and alloy-specific modeling for evaluating aluminum recycling strategies for vehicles. JOM 66, 2262–2271 (2014)

### This study discusses alloy design challenges and solution approaches to make better recycling use of mixed shredded aluminium scrap from end-of-life vehicles.

- Enkvist, P. A. Klevnas, P. The Circular Economy—A Powerful Force for Climate Mitigation. (Material Economics Sverige, 2018).
- Haupt, M., Vadenbo, C., Zeltner, C. & Hellweg, S. Influence of input-scrap quality on the environmental impact of secondary steel production. *J. Ind. Ecol.* 21, 391–401 (2017)
- Daehn, K. E., Cabrera Serrenho, A. & Allwood, J. M. How will copper contamination constrain future global steel recycling? *Environ. Sci. Technol.* 51, 6599–6606 (2017)

## This paper focuses on copper contamination in steel through a detailed characterization of copper in the global steel system along the steel supply chain to quantify the maximum concentration of copper in steel products.

- Elshkaki, A., Reck, B. K. & Graedel, T. E. Anthropogenic nickel supply. demand, and associated energy and water use. Resour. Conserv. Recycling 125, 300–307 (2017).
- Reck, B. K. & Rotter, V. S. Comparing growth rates of nickel and stainless steel use in the early 2000s. J. Ind. Ecol. 16, 518–528 (2012).
- Worrell, E. & Reuter, M. A. Handbook of Recycling: State-of-the-art for Practitioners, Analysts, and Scientists (Elsevier, 2014).

# This is a comprehensive presentation of materials recovery that provides insight into recycling technologies, economics, and policy for a comprehensive set of materials, focusing on steel, copper and nickel (in the context of stainless steel).

- Takeda, O. & Okabe, T. H. Current status of titanium recycling and related technologies. JOM 71, 1981–1990 (2019).
- Pogatscher, S. et al. Statistical and thermodynamic optimization of trace-element modified Al-Mg-Si-Cu alloys. In *Light Metals 2015* (ed. Hyland, M.) 263–270 (2015).
- Liu, Z.-K. & Wang, Y. Computational Thermodynamics of Materials (Cambridge Univ. Press, 2016).
- Ebenberger, P. et al. Processing-controlled suppression of Lüders elongation in AlMgMn alloys. Scr. Mater. 166, 64–67 (2019).
- Koch, G. H., Brongers, M. P. H., Thompson, N. G., Virmani, Y. P. & Payer, J. H. Corrosion Costs and Preventive Strategies in the United States. Technical report 200224 (Federal Highway Administration, 2002).
- Song, R. G. et al. Stress corrosion cracking and hydrogen embrittlement of an Al-Zn-Mg-Cu alloy. Acta Mater. 52, 4727-4743 (2004).

- Martin, M. L., Dadfarnia, M., Nagao, A., Wang, S. & Sofronis, P. Enumeration of the hydrogen-enhanced localized plasticity mechanism for hydrogen embrittlement in structural materials. Acta Mater. 165, 734-750 (2019).
  - This is an overview of experiments and simulation results about the hydrogenenhanced localized plasticity (HELP) mechanism as a viable hydrogen embrittlement mechanism in structural materials for a hydrogen economy.
- Odette, G. R., Alinger, M. J. & Wirth, B. D. Recent developments in irradiation-resistant steels, Annu. Rev. Mater. Res. 38, 471-503 (2008).
- 54. Zinkle, S. J. & Busby, J. T. Structural materials for fission & fusion energy. Mater. Today 12,
- Yao, Y., Wei, J., Liu, J., Wang, Z. & Wang, Y. Thermal ageing embrittlement of casting duplex stainless steels for nuclear power plant. In 18th Int. Conf. on Nuclear Engineering Vol. 5, 625-630 (2011).
- 56. Li. S., Wang, Y. & Wang, X. Effects of ferrite content on the mechanical properties of thermal aged duplex stainless steels, Mater, Sci. Eng. A 625, 186-193 (2015).
- 57. Ritchie, R. O. Mechanisms of fatigue crack propagation in metals, ceramics and composites: role of crack tip shielding. Mater. Sci. Eng. 103, 15-28 (1988).
- 58. Koyama, M, et al, Bone-like crack resistance in hierarchical metastable nanolaminate steels, Science 355, 1055-1057 (2017).
- 59. Dai, K. & Shaw, L. Analysis of fatigue resistance improvements via surface severe plastic deformation. Int. J. Fatigue 30, 1398-1408 (2008).
- 60. Baxevanis, T., Parrinello, A. F. & Lagoudas, D. C. On the fracture toughness enhancement due to stress-induced phase transformation in shape memory alloys. Int. J. Plast. 50, 158-169 (2013).
- Pippan, R. & Hohenwarter, A. Fatigue crack closure: a review of the physical phenomena. 61. Fatigue Fract. Eng. Mater. Struct. 40, 471-495 (2017).
- Ritchie, R. O. Mechanisms of fatigue-crack propagation in ductile and brittle solids. Int. J. 62 Fract. 100, 55-83 (1999).
- 63. Zhang, S. et al. Autonomous filling of grain-boundary cavities during creep loading in Fe-Mo alloys. Metall. Mater. Trans. A 47, 4831-4844 (2016).
- Takahashi, J., Kawakami, K., Kobayashi, Y. & Tarui, T. The first direct observation of hydrogen trapping sites in TiC precipitation-hardening steel through atom probe tomography. Scr. Mater. 63, 261-264 (2010).
- Chen, Y. S. et al. Direct observation of individual hydrogen atoms at trapping sites in a ferritic steel. Science 355, 1196-1199 (2017).
- 66. Chang, Y. et al. Quantification of solute deuterium in titanium deuteride by atom probe tomography with both laser pulsing and high-voltage pulsing: influence of the surface electric field. New J. Phys. 21, 053025 (2019).
- 67. Kirchheim, R. Solid solutions of hydrogen in complex materials. In Solid State Phys Vol. 59 (eds Ehrenreich, H. & Spaepen, F.) 203-291 (2004).
- 68. Pundt, A. & Kirchheim, R. Hydrogen in metals: microstructural aspects. Annu. Rev. Mater. Res. 36, 555-608 (2006).
- 69. Nozaka, K., Shield, C. K. & Hajiar, J. F. Effective bond length of carbon-fiber-reinforced polymer strips bonded to fatigued steel bridge I-girders. J. Bridge Eng. 10, 195-205 (2005).
- 70 Garlock, M., Paya-Zaforteza, I., Kodur, V. & Gu, L. Fire hazard in bridges: review, assessment and repair strategies. Eng. Struct. 35, 89-98 (2012).
- 71. Yilmaz, O., Gindy, N. & Gao, J. A repair and overhaul methodology for aeroengine components. Robot. Comput.-Integr. Manuf. 26, 190-201 (2010).
- 72 Mokadem, S., Bezençon, C., Hauert, A., Jacot, A. & Kurz, W. Laser repair of superalloy single crystals with varying substrate orientations. Metall. Mater. Trans. A 38, 1500–1510 (2007)
- Grabowski, B. & Tasan, C. C. Self-healing metals. Adv. Polym. Sci. 273, 387-407 (2016).
- Hager, M. D., Greil, P., Leyens, C., Van Der Zwaag, S. & Schubert, U. S. Self-healing materials. Adv. Mater. 22, 5424-5430 (2010).
- Hautakangas, S., Schut, H., van der Zwaag, S. & van Dijk, N. H. The role of the aging temperature on the self-healing kinetics in an underaged Aa2024 aluminium alloy. Proc. 1st Int. Conf. on Self Healing Materials 1–7 (2007).
- 76. He, S. M., Van Dijk, N. H., Schut, H., Peekstok, E. R. & Van Der Zwaag, S. Thermally activated precipitation at deformation-induced defects in Fe-Cu and Fe-Cu-B-N alloys studied by positron annihilation spectroscopy. Phys. Rev. B 81, 094103 (2010).
- Zhang, S. et al. Autonomous repair mechanism of creep damage in Fe-Au and Fe-Au-B-N alloys. Metall. Mater. Trans. A 46, 5656-5670 (2015).
- 78. Hoefnagels, J. P. M., Tasan, C. C., Maresca, F., Peters, F. J. & Kouznetsova, V. G. Retardation of plastic instability via damage-enabled microstrain delocalization. J. Mater. Sci. 50. 6882-6897 (2015).
- Mei. Z. & Morris, J. W. Analysis of transformation-induced crack closure. Eng. Fract. Mech. 79. **39**, 569-573 (1991).
- 80 Li, S. L. et al. Annealing induced recovery of long-term thermal aging embrittlement in a duplex stainless steel. Mater. Sci. Eng. A 564, 85-91 (2013).
- 81 Hoefnagels, J. P. M., Du, C. & Tasan, C. C. Laser-induced toughening inhibits cut-edge failure in multi-phase steel. Scr. Mater. (in the press).
- Cooper, D. R. & Allwood, J. M. Reusing steel and aluminum components at end of product life. Environ. Sci. Technol. 46, 10334-10340 (2012).
- Roters, F. et al. DAMASK The Düsseldorf Advanced Material Simulation Kit for modeling 83. multi-physics crystal plasticity, thermal, and damage phenomena from the single crystal up to the component scale. Comput. Mater. Sci. 158, 420-478 (2019).
- 84. Andersson, J. O., Helander, T., Höglund, L., Shi, P. & Sundman, B. Thermo-Calc & DICTRA, computational tools for materials science. Calphad 26, 273-312 (2002).
- Kalidindi, S. R. Hierarchical Materials Informatics (Butterworth-Heinemann, 2015)
- Kalidindi, S. R. Materials, data, and informatics. In Hierarchical Materials Informatics 1–32 (Butterworth-Heineman, 2015).
- Kim, E. et al. Materials synthesis insights from scientific literature via text extraction and machine learning. Chem. Mater. 29, 9436-9444 (2017).
- Kim, E., Huang, K., Jegelka, S. & Olivetti, E. A. Virtual screening of inorganic materials synthesis parameters with deep learning. npj Comput. Mater. 3, 53 (2017).

- Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. npj Comput. Mater. 2, 16028 (2016).
- Cameron, B. C. & Tasan, C. C. Microstructural damage sensitivity prediction using spatial 90. statistics. Sci. Rep. 9, 2774 (2019).
- Densley Tingley, D., Cooper, S. & Cullen, J. Understanding and overcoming the barriers to structural steel reuse, a UK perspective. J. Clean. Prod. 148, 642-652 (2017).
- Edmonds, D. V. et al. Quenching and partitioning martensite: a novel steel heat treatment. Mater. Sci. Eng. A 438-440, 25-34 (2006).
- Grässel, O., Krüger, L., Frommeyer, G. & Meyer, L. W. High strength Fe-Mn-(Al, Si) TRIP/TWIP steels development-properties-application. Int. J. Plast. 16, 1391-1409 (2000)

#### The work introduced one of the clearest examples of how transformation induced plasticity (TRIP) or twinning induced plasticity (TWIP) effects can improve property combinations towards reducing the density and increasing the strength of steels.

- 94. Tasan, C. C. et al. An overview of dual-phase steels: advances in microstructure. oriented processing and micromechanically guided design, Annu. Rev. Mater. Res. 45. 391-431 (2015).
- Galán, J., Samek, L., Verleysen, P., Verbeken, K. & Houbaert, Y. Advanced high strength 95 steels for automotive industry. Rev. Metal. 48, 118-131 (2012).
- 96 Heinz, A. et al. Recent development in aluminium alloys for aerospace applications. Mater. Sci. Eng. A 280, 102-107 (2000).
- 97 Williams, J. C. & Starke, E. A. Progress in structural materials for aerospace systems. Acta Mater. 51, 5775-5799 (2003).
- 98 Boyer, R. R. An overview on the use of titanium in the aerospace industry, Mater. Sci. Eng. A 213, 103-114 (1996).
- Lee, W. G., Kim, J. S., Sun, S. J. & Lim, J. Y. The next generation material for lightweight railway car body structures: magnesium alloys. Proc. Inst. Mech. Eng. F 232, 25-42 (2018).
- 100. Helms, H. & Lambrecht, U. The potential contribution of light-weighting to reduce transport energy consumption. Int. J. Life Cycle Assess. 12, 58-64 (2007).
- 101. Kim, S.-H., Kim, H. & Kim, N. J. Brittle intermetallic compound makes ultrastrong lowdensity steel with large ductility. Nature 518, 77-79 (2015).
- 102. Li, Z., Pradeep, K. G., Deng, Y., Raabe, D. & Tasan, C. C. Metastable high-entropy dualphase alloys overcome the strength-ductility trade-off. Nature 534, 227–230 (2016).
- 103. Lei, Z. et al. Enhanced strength and ductility in a high-entropy alloy via ordered oxygen complexes. Nature 563, 546-550 (2018).
- 104. Wang, Y., Chen, M., Zhou, F. & Ma, E. High tensile ductility in a nanostructured metal. Nature 419, 912-915 (2002).
- 105. Matlock, D. K. & Speer, J. G. Third generation of AHSS: microstructure design concepts. In Microstructure and Texture in Steels (eds Haldar, A. Suwas, S. & Bhattacharjee, D.) 185-205 (2009).

#### The work takes an overview of the increased interest in the development of thirdgeneration advanced high-strength steels.

- 106. Wang, M. M., Tasan, C. C., Ponge, D., Dippel, A. C. & Raabe, D. Nanolaminate transformation-induced plasticity-twinning-induced plasticity steel with dynamic strain partitioning and enhanced damage resistance. Acta Mater. 85, 216-228 (2015).
- 107. Fischer, F. D. et al. New view on transformation induced plasticity (TRIP). Int. J. Plast. 16, 723-748 (2000).
- 108. Steinmetz, D. R., Ja, T., Wietbrock, B. & Eisenlohr, P. Revealing the strain-hardening behavior of twinning-induced plasticity steels: theory, simulations, experiments. Acta. Mater. 61, 494-510 (2013).
- 109. Chen, L., Zhao, Y. & Qin, X. Some aspects of high manganese twinning-induced plasticity (TWIP) steel, a review. Acta Metall. Sin. 26, 1-15 (2013).
- Güvenç, O., Roters, F., Hickel, T. & Bambach, M. ICME for crashworthiness of TWIP steels: from ab initio to the crash performance. JOM 67, 120-128 (2015).
- 111. Sun, F. et al. Deformation microstructure and mechanisms in a metastable β titanium alloy exhibiting TWIP and TRIP effects. Mater. Sci. Forum 783-786, 1360-1365 (2014).
- 112. Hong, D. H., Lee, T. W., Lim, S. H., Kim, W. Y. & Hwang, S. K. Stress-induced hexagonal close-packed to face-centered cubic phase transformation in commercialpurity titanium under cryogenic plane-strain compression. Scr. Mater. 69, 405-408 (2013)
- 113. Springer, H., Baron, C., Szczepaniak, A., Uhlenwinkel, V. & Raabe, D. Stiff, light, strong and ductile: nano-structured high modulus steel, Sci. Rep. 7, 2757 (2017).
- 114. Miller, W. S. et al. Recent development in aluminium alloys for the automotive industry. Mater. Sci. Eng. A 280, 37-49 (2000).

#### This seminal contribution provides a critical overview of the use of aluminium as a structural material in cast, extruded or sheet conditions.

- 115. Mordike, B. L. & Ebert, T. Magnesium properties—applications—potential. *Mater. Sci. Eng.* A 302, 37-45 (2001).
- Marlaud, T., Deschamps, A., Bley, F., Lefebvre, W. & Baroux, B. Influence of alloy composition and heat treatment on precipitate composition in Al-Zn-Mg-Cu alloys. Acta Mater. 58, 248-260 (2010).
- Starke, E. A., Sanders, T. H. & Palmer, I. G. New approaches to alloy development in the Al-Li System. JOM 33, 24-33 (1981).
- Suzuki, A., Saddock, N. D., Jones, J. W. & Pollock, T. M. Solidification paths and eutectic intermetallic phases in Mg-Al-Ca ternary alloys. Acta Mater. 53, 2823-2834 (2005).
- Sandlöbes, S. et al. A rare-earth free magnesium alloy with improved intrinsic ductility. Sci. Rep. 7, 10458 (2017).
- 120. Sandlöbes, S., Zaefferer, S., Schestakow, I., Yi, S. & Gonzalez-Martinez, R. On the role of non-basal deformation mechanisms for the ductility of Mg and Mg-Y alloys. Acta Mater. 59, 429-439 (2011).
- 121. Counts, W. A., Friák, M., Raabe, D. & Neugebauer, J. Using ab initio calculations in designing bcc MgLi-X alloys for ultra-lightweight applications. In Advanced Engineering Materials Vol. 12 (ed. Quandt, E.) 1198-1205 (Pergamon, 2010).

#### Review

- Zhu, L., Li, N. & Childs, P. R. N. Light-weighting in aerospace component and system design. *Propuls. Power Res.* 7, 103–119 (2018).
- Kalidindi, S. R. & De Graef, M. Materials data science: current status and future outlook. Annu. Rev. Mater. Res. 45, 171–193 (2015).
- Shao, C. W. et al. Improvement of low-cycle fatigue resistance in TWIP steel by regulating the grain size and distribution. Acta Mater. 134, 128–142 (2017).
- Niendorf, T. et al. Functionally graded alloys obtained by additive manufacturing. Adv. Eng. Mater. 16, 857–861 (2014).
- Xu, W. et al. Ti-6Al-4V additively manufactured by selective laser melting with superior mechanical properties. JOM 67, 668–673 (2015).
- Long, J. et al. Improved fatigue resistance of gradient nanograined Cu. Acta Mater. 166, 56–66 (2019)
- Terada, D., Houda, H. & Tsuji, N. Effect of grain size distribution on mechanical properties of ultrafine grained Al severely deformed by ARB process and subsequently annealed. J. Phys. Conf. Ser. 240, 012111 (2010).
- Kim, S. I. et al. Dense dislocation arrays embedded in grain boundaries for highperformance bulk thermoelectrics. Science 348, 109–114 (2015).
- Kim, S. W., Li, X., Gao, H. & Kumar, S. In situ observations of crack arrest and bridging by nanoscale twins in copper thin films. Acta Mater. 60, 2959–2972 (2012).
- Suresh, S. Graded materials for resistance to contact deformation and damage. Science 292, 2447–2451 (2001).
- Ho, S. & Lavernia, E. J. Thermal residual stresses in functionally graded and layered 6061 Al/SiC materials. Metall. Mater. Trans. A 27, 3241–3249 (1996).
- Carpenter, J. S., McCabe, R. J., Mayeur, J. R., Mara, N. A. & Beyerlein, I. J. Interface-driven plasticity: the presence of an interface affected zone in metallic lamellar composites. Adv. Eng. Mater. 17, 109–114 (2015).
- Akasheh, F., Zbib, H. M., Hirth, J. P., Hoagland, R. G. & Misra, A. Dislocation dynamics analysis of dislocation intersections in nanoscale metallic multilayered composites. J. Appl. Phys. 101, 084314 (2007).
- Knorr, I., Cordero, N. M., Lilleodden, E. T. & Volkert, C. A. Mechanical behavior of nanoscale Cu/PdSi multilayers. Acta Mater. 61, 4984–4995 (2013).
- Ma, Z. et al. Strength gradient enhances fatigue resistance of steels. Sci. Rep. 6, 22156 (2016)
- Mozaffari, A., Danesh Manesh, H. & Janghorban, K. Evaluation of mechanical properties and structure of multilayered Al/Ni composites produced by accumulative roll bonding (ARB) process. J. Alloys Compd. 489, 103–109 (2010).
- Tayyebi, M. & Eghbali, B. Study on the microstructure and mechanical properties of multilayer Cu/Ni composite processed by accumulative roll bonding. *Mater. Sci. Eng. A* 559, 759–764 (2013).
- Salomon, S. et al. Combinatorial synthesis and high-throughput characterization of the thin film materials system Co–Mn–Ge: composition, structure, and magnetic properties. *Phys. Status Solidi A* 212, 1969–1974 (2015).
- König, D., Eberling, C., Kieschnick, M., Virtanen, S. & Ludwig, A. High-throughput investigation of the oxidation and phase constitution of thin-film Ni–Al–Cr materials libraries. Adv. Eng. Mater. 17, 1365–1373 (2015).

- Carroll, B. E. et al. Functionally graded material of 304L stainless steel and Inconel 625 fabricated by directed energy deposition: characterization and thermodynamic modeling. Acta Mater. 108, 46–54 (2016).
- Yin, S. et al. Hybrid additive manufacturing of Al-Ti6Al4V functionally graded materials with selective laser melting and cold spraying. J. Mater. Process. Technol. 255, 650–655 (2018).
- Muller, P., Hascoet, J. Y. & Mognol, P. Toolpaths for additive manufacturing of functionally graded materials (FGM) parts. *Rapid Prototyping J.* 20, 511–522 (2014).
- Hofmann, D. C. et al. Developing gradient metal alloys through radial deposition additive manufacturing. Sci. Rep. 4, 5357 (2014).
- 145. Li, W. et al. Fabrication and characterization of a functionally graded material from Ti-6Al-4V to SS316 by laser metal deposition. *Addit. Manuf.* **14**, 95-104 (2017)
- Knoll, H. et al. Combinatorial alloy design by laser additive manufacturing. Steel Res. Int. 88, 1600416 (2017).
- Robbins, J., Owen, S. J., Clark, B. W. & Voth, T. E. An efficient and scalable approach for generating topologically optimized cellular structures for additive manufacturing. Addit. Manuf. 12, 296–304 (2016).
- Murphy, C. (ed.) McKinsey Sustainability & Resource Productivity Practice https://assets. mckinsey.com/-/media/FF702F8557624AD0884F116174A60298.ashx (McKinsey Global Institute, 2014).
- de Pee, A. et al. Decarbonization of Industrial Sectors: The Next Frontier (McKinsey, 2018).
- 150. International Aluminium Association. Global aluminium cycle 2017. World Aluminium www.world-aluminium.org/statistics/massflow (2017).
- Pauliuk, S., Milford, R. L., Müller, D. B. & Allwood, J. M. The steel scrap age. Environ. Sci. Technol. 47, 3448–3454 (2013).
- Duke, R. Policy Options to Deeply Decarbonize American Industry. Discussion Draft for July 18, 2019 Brooking Institution Workshop (2019).

Acknowledgements E.A.O. acknowledges support from B. Reck in quantifying nickel flows.

Author contributions All authors contributed equally to this paper.

Competing interests The authors declare no competing interests.

#### Additional information

Correspondence and requests for materials should be addressed to D.R., C.C.T. or E.A.O. Peer review information Nature thanks Daniel Cooper, Elizabeth Holm, Gabriella Tranell and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Reprints and permissions information is available at http://www.nature.com/reprints.

© Springer Nature Limited 2019

# Recycling lithium-ion batteries from electric vehicles

https://doi.org/10.1038/s41586-019-1682-5

Received: 14 January 2019

Accepted: 23 July 2019

Published online: 6 November 2019

Gavin Harper<sup>1,2,3</sup>\*, Roberto Sommerville<sup>1,2,4</sup>, Emma Kendrick<sup>1,2,3</sup>, Laura Driscoll<sup>1,2,5</sup>, Peter Slater<sup>1,2,5</sup>, Rustam Stolkin<sup>1,2,3,6</sup>, Allan Walton<sup>1,2,3</sup>, Paul Christensen<sup>1,7</sup>, Oliver Heidrich<sup>1,7,8</sup>, Simon Lambert<sup>1,7</sup>, Andrew Abbott<sup>1,9</sup>, Karl Ryder<sup>1,9</sup>, Linda Gaines<sup>10</sup> & Paul Anderson<sup>1,2,5</sup>\*

Rapid growth in the market for electric vehicles is imperative, to meet global targets for reducing greenhouse gas emissions, to improve air quality in urban centres and to meet the needs of consumers, with whom electric vehicles are increasingly popular. However, growing numbers of electric vehicles present a serious waste-management challenge for recyclers at end-of-life. Nevertheless, spent batteries may also present an opportunity as manufacturers require access to strategic elements and critical materials for key components in electric-vehicle manufacture: recycled lithium-ion batteries from electric vehicles could provide a valuable secondary source of materials. Here we outline and evaluate the current range of approaches to electric-vehicle lithium-ion battery recycling and re-use, and highlight areas for future progress.



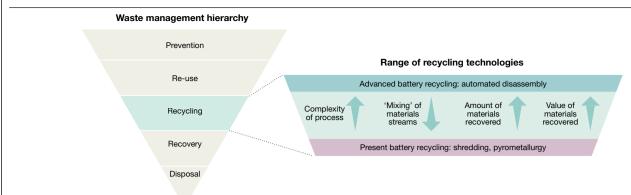
The electric-vehicle revolution, driven by the imperatives to decarbonize personal transportation in order to meet global targets for reductions in greenhouse gas emissions and improve air quality in urban centres, is set to change the automotive industry radically. In 2017, sales of electric vehicles exceeded one million cars per year worldwide for the first time<sup>1</sup>. Making conservative assumptions of an average battery pack weight of 250 kg and volume of half a cubic metre, the resultant pack wastes would comprise around 250,000 tonnes and half a million cubic metres of unprocessed pack waste, when these vehicles reach the end of their lives. Although re-use and current recycling processes can divert some of these wastes from landfill, the cumulative burden of electric-vehicle waste is substantial given the growth trajectory of the electric-vehicle market. This waste presents a number of serious challenges of scale; in terms of storing batteries before repurposing or final disposal, in the manual testing and dismantling processes required for either, and in the chemical separation processes that recycling entails.

Given that the environmental footprint of manufacturing electric vehicles is heavily affected by the extraction of raw materials and production of lithium ion batteries, the resulting waste streams will inevitably place different demands on end-of-life dismantling and recycling systems. In the waste management hierarchy, re-use is considered preferable to recycling (Fig. 1). Because considerable value is embedded in manufactured lithium-ion batteries (LIBs), it has been suggested that their use should be cascaded through a hierarchy of applications to optimize material use and life-cycle impacts<sup>2</sup>. Markets for energy storage are under development as energy regulators in various locations transition to cleaner energy sources. Energy storage is particularly soughtafter in areas where weak grids require reinforcement, where high penetration of renewables requires supply to be balanced with demand, where there is an opportunity for trading energy with the grid and in offgrid applications. Second-use battery projects have started to develop in locations where there is regulatory and market alignment. However, large concentrations of waste-be it for refurbishment, re-manufacture, dismantling or final disposal—can create substantial challenges. A fire in stockpiled tyres in Powys, Wales, for example, smouldered for fifteen years from 1989 to 2004. Since the electrode materials in LIBs are far more reactive than tyre rubber<sup>3</sup>, without a proactive and economically sound waste-management strategy for LIBs there are potentially greater dangers associated with stockpiling of end-of-life LIBs. Already the number of fires being reported in metal-recovery facilities is increasing<sup>4</sup>, owing to the illicit or accidental concealment of (consumer) LIBs in the guise of, for example, lead-acid batteries. Among examples of recent major fires are those that took place in metal-recovery facilities in Shoreway, San Carlos, USA, in September 2016<sup>5</sup>, Guernsey in August 2018 and Tacoma, Washington, USA, in September 2018.

Waste may also represent a valuable resource. Elements and materials contained in electric-vehicle batteries are not available in many nations and access to resources is crucial in ensuring a stable supply chain. In the future, electric vehicles may prove to be a valuable secondary resource for critical materials, and it has been argued that high-cobalt-content batteries should be recycled immediately to bolster cobalt supplies<sup>6</sup>. If tens of millions of electric vehicles are to be produced annually, careful husbandry of the resources consumed by electric-vehicle battery manufacturing will surely be essential to ensure the sustainability of the automotive industry of the future, as will a material- and energyefficient 3R system (reduce, re-use, recycle). Here we give an overview

Faraday Institution, ReLiB Project, University of Birmingham, Birmingham, UK. 2Birmingham Centre for Strategic Elements and Critical Materials, University of Birmingham, Birmingham, UK. 3School of Metallurgy and Materials, University of Birmingham, Birmingham, UK. 4School of Chemical Engineering, University of Birmingham, Birmingham, UK. 5School of Chemistry, University of Birmingham, Birmingham, UK. 6National Centre for Nuclear Robotics, University of Birmingham, Birmingham, UK. 7School of Engineering, Newcastle University, Newcastle, UK. <sup>8</sup>Tyndall Centre for Climate Change Research, Newcastle University, Newcastle, UK. <sup>9</sup>Materials Centre, University of Leicester, Leicester, UK. <sup>10</sup>ReCell Center, Argonne National Laboratory, Lemont, IL, USA, \*e-mail: q.d.i.harper@bham.ac.uk: p.a.anderson@bham.ac.uk

#### Review



**Fig. 1**| The waste management hierarchy and range of recycling options. The waste management hierarchy is a concept that was developed from the Council Directive 75/442/EEC of 15 July 1975 (https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A31975L0442) on waste by the Dutch politician Ad Lansink, in 1979, who presented to the Dutch parliament a simple schematic representation that has been termed 'Lansink's Ladder', ranking waste management options from the most to least environmentally desirable options. Here, that hierarchy is expanded to consider the range of battery recycling technologies. 'Prevention' means that LIBs are designed to use less-critical

of the current state of the art and identify some of the important issues relating to the end-of-life management of electric-vehicle LIBs.

#### Social and environmental impacts of LIBs

If we consider the two main modes of primary production, it takes 250 tons of the mineral ore spodumene<sup>7,8</sup> when mined, or 750 tons of mineral-rich brine<sup>7,8</sup> to produce one ton of lithium. The processing of large amounts of raw materials can result in considerable environmental impacts<sup>9</sup>. Production from brine, for example, entails drilling a hole in the salt flat, and pumping of the mineral-rich solution to the surface. However, this mining activity depletes water tables. In Chile's Salar de Atacama, a major centre of lithium production, 65% of the region's water is consumed by mining activities<sup>9</sup>. This affects farmers in the region who must then import water from other regions. The demands on water from the processing of lithium produced in this way are substantial, with a ton of lithium requiring 1,900 tons of water to extract, which is consumed by evaporation<sup>9</sup>.

By contrast, secondary production would require only 28 tons of used LIBs<sup>7,8,10</sup> (around 256 used electric-vehicle LiBs<sup>8</sup>). The net impact of LIB production can be greatly reduced if more materials can be recovered from end-of-life LIBs, in as close to usable form as possible<sup>11</sup>. However, in the rapid-growth phase of the electric-vehicle market, recycling alone cannot come close to replenishing mineral supplies<sup>12</sup>. LIBs are anticipated to last 15–20 years<sup>12</sup> based on calendar aging (the aging due to time since manufacture) predictions—three times longer than lead—acid batteries<sup>12</sup>. Initial concerns regarding resource constraints for LIB production scale-up focused on lithium<sup>13</sup>; however, in the near term, reserves of lithium are unlikely to present a constraint<sup>14,15</sup>.

Of greater immediate concern are cobalt reserves<sup>16</sup>, which are geographically concentrated (mainly in the politically unstable Democratic Republic of the Congo). These have experienced wild short-term price fluctuations and raise multifarious social, ethical and environmental concerns around their extraction, including artisanal mines employing child labour<sup>17</sup>. In addition to the environmental imperative for recycling, there are clearly serious ethical concerns with the materials supply chain, and these social burdens are borne by some of the world's most vulnerable people. Given the global nature of the industry, this will require international coordination to support a concerted push towards recycling LIBs and a circular economy in materials<sup>18</sup>.

materials (high economic importance, but at risk of short supply) and that electric vehicles should be lighter and have smaller batteries. 'Re-use' means that electric-vehicle batteries should have a second use. 'Recycling' means that batteries should be recycled, recovering as much material as possible and preserving any structural value and quality (for example, preventing contamination). 'Recovery' means using some battery materials as energy for processes such as fuel for pyrometallurgy. Finally, 'disposal' means that no value is recovered and the waste goes to landfill.

#### **Battery assessment and disassembly**

The waste-management hierarchy considers re-use to be preferable to recycling (Fig. 1). As considerable value is embedded in manufactured LIBs, it has been suggested that their use should be cascaded through a hierarchy of applications to optimize material use and life-cycle impacts². Energy stored over energy invested (ESOI)—the ratio between the energy that must be invested into manufacturing the battery and the electrical energy that it will store over its useful life—is a metric used to compare the efficacy of different energy-storage technologies. Clearly, ESOI figures will improve if end-of-life electric-vehicle batteries can be used in second-use applications for which the battery performance is less critical.

Profitable second-use applications also provide a potential value stream that can offset the eventual cost of recycling, and already a healthy market is developing in used electric-vehicle batteries for energy storage in certain localities, with demand potentially outstripping supply. For the moment the economics of the decision whether to recycle or re-use are set firmly in favour of re-use. The main factors are (1) the refurbishment cost of putting the battery into a second-use application and (2) any credit that would accrue as the result of recycling the battery instead; if the second-use price were to fall below the sum of the refurbishment cost and the recycling credit, then recycling would be the economically favoured option<sup>19</sup>. In time, it is anticipated<sup>19</sup> that the supply of used electric-vehicle batteries will far exceed the quantity that the second-use market can absorb. It must be remembered, therefore, that—if disposal to landfill is to be avoided—recycling must be the ultimate fate of all LIBs, even if they first have a second use.

Given that stockpiling of waste batteries is potentially unsafe and environmentally undesirable, if direct re-use of an LIB module is not possible, it must be repaired or recycled. End-of-life LIB recycling could provide important economic benefits, avoiding the need for new mineral extraction<sup>20</sup> and providing resilience against vulnerable links<sup>21</sup> and supply risks<sup>22</sup> in the LIB supply chain. For most remanufacture and recycling processes, battery packs must be disassembled to module level at least. However, the hazards associated with battery disassembly are also numerous<sup>23,24</sup>. Disassembly of battery packs from automotive applications requires high-voltage training and insulated tools to prevent electrocution of operators or short-circuiting of the pack. Short-circuiting results in rapid discharge, which may lead to heating and thermal runaway. Thermal runaway may result in the generation of particularly noxious byproducts, including HF gas<sup>25</sup>, which along with

other product gases may become trapped and ultimately result in cells exploding<sup>23</sup>. The cells also present a chemical hazard owing to the flammable electrolyte, toxic and carcinogenic electrolyte additives, and the potentially toxic or carcinogenic electrode materials.

#### Diagnostics of battery pack, modules and cells

'State of health' is the degree to which a battery meets its initial design specifications. Over time as the battery degrades, its performance varies from its initial condition. The units are percentage points, with 100% indicating a state of health that is identical to that of a new battery meeting its design specification. (Some new batteries may leave the factory deviating from design specifications, and having less than 100% state of health.) The 'state of charge' is the degree to which a battery is charged or discharged. Again, the units are percentage points, with 0% indicating empty and 100% indicating full).

Battery repurposing—the re-use of packs, modules and cells in other applications such as charging stations and stationary energy storagerequires accurate assessment of both the state of health, to categorize whether batteries are best suited for re-use (and if so, for which applications), remanufacture or recycling, and the state of charge, for safety reasons in some recycling processes. For high-throughput triage and gateway testing of batteries at scale, the optimal approach involves in situ techniques for monitoring cells in service to enable advance warning of possible cell replacement, and module or pack reconditioning, rather than complete repurposing at a low level of state of health owing to a few failing cells.

Electrochemical impedance spectroscopy can give information on the state of health of cells, modules and, potentially, full packs<sup>26</sup>, and also an indication of aging mechanisms such as lithium plating. Such measurements have the potential to inform a decision matrix for re-use or disassembly and processing and, importantly, to identify potential hazards that would have further consequences for downstream processing. Electrochemical impedance spectroscopy has been researched for gateway testing in primary production, for example, in a large battery production plant in the UK<sup>27,28</sup>. A number of electric-vehicle manufacturers plan to use similar technologies to manage and maintain electricvehicle battery packs through the identification and replacement of failing modules in the field. Substantial advantages in cost, safety and throughput time are anticipated if this process can be mostly or fully automated<sup>27,29</sup>. In future, more advanced diagnostic functionality will be embedded in battery management systems, providing data that can be interrogated at end-of-life.

#### Challenges of pack and module disassembly

Different vehicle manufacturers have adopted different approaches for powering their vehicles, and electric vehicles on the market possess a wide variety of different physical configurations, cell types and cell chemistries. This presents a challenge for battery recycling. Figure 2 details three different types of battery cell design, and their respective packs from electric vehicles in the marketplace from model year 2014. It can be seen that the three vehicles possess very different physical configurations, requiring different approaches for disassembly, particularly regarding automation. It can be seen in Fig. 2 that at the different scales of disassembly, the format and relative size of the different components differ, presenting challenges for automation. The differing form factors and capacities may also restrict applications for re-use. And finally, Fig. 2 illustrates that manufacturers employ varying cell chemistries (see Fig. 3), which will necessitate different approaches to materials reclamation and strongly affect the overall economics of recycling. Whereas the prismatic and pouch cells have planar electrodes, the cylindrical cells are tightly coiled, presenting additional challenges to separating the electrodes for direct recycling processes.

For repurposing and second-use applications, automotive battery packs are currently dismantled by hand for either the second use of the modules or for recycling. The weights and high voltages of traction batteries mean that qualified employees and specialized tools are required for such dismantling<sup>25</sup>. This is a challenge for an industry in transition with a shortage of skills. An Institute of the Motor Industry survey found only 1.000 trained technicians in the UK capable of servicing electric vehicles<sup>30</sup>, with another 1,000 in training. Given there are 170,000 motor technicians in the UK, this represents less than 2% of the workforce. There is concern that untrained mechanics may risk their lives repairing electric vehicles<sup>31</sup>, and these concerns logically extend to those handling vehicles at the end-of-life. Additionally, it has been suggested32 that manual dismantling, in countries with high labour costs, is uneconomic with respect to revenues from extracted materials or components. Vehicle design has to strike compromises between crash safety, centre of gravity and space optimization, which must be balanced against serviceability<sup>25</sup>. These conflicting design objectives often result in designs that are not optimized for recyclability, and that can be time-consuming to disassemble manually<sup>25</sup>.

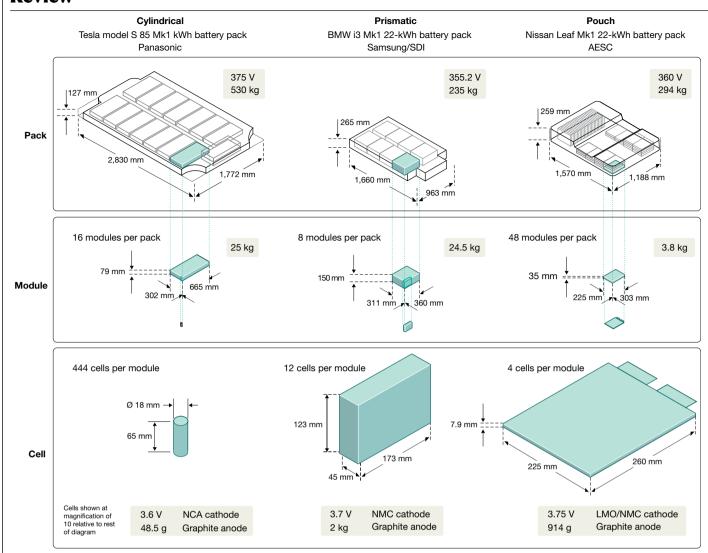
#### **Automating battery disassembly**

Robotic battery disassembly could eliminate the risk of harm to human workers, and increased automation would reduce cost, potentially making recycling economically viable. This is being piloted in a number of current research projects<sup>33–36</sup>. Importantly, automation could also improve the mechanical separation of materials and components, enhancing the purity of segregated materials and making downstream separation and recycling processes more efficient. The automation of the dismantling of automotive batteries, however, presents major challenges. This is because robotics and automation in the manufacturing sector rely on highly structured environments, in which robots make preprogrammed repetitive actions with respect to exactly known objects in fixed positions. In contrast, the development of robotic systems that can generalize to a variety of objects, and handle uncertainty, remains a major challenge at the frontier of artificial intelligence research. It is important to consider the complexity of vehicle battery disassembly from this perspective.

At present there is no standardization<sup>37</sup> of design for battery packs, modules or cells within the automotive sector, and it is unlikely that this will happen in the near future. Other battery-reliant products, such as mobile phones, have seen an exponential proliferation of different sizes, shapes and types of battery over the past two decades. At present, much of the factory assembly of these batteries is done by human workers and remains unautomated. Their disassembly and waste-handling typically involve even less structured environments, with much greater uncertainties, than a manufacturing assembly line.

Nevertheless, some progress has been made towards automated sorting of consumer batteries. The Optisort system  $^{38,39}$  uses computer vision algorithms to recognize the labels on batteries, and then pneumatic actuators to segregate batteries into different bins according to their type of chemistry. However, Optisort is currently limited to AA and AAA batteries, and a large amount of pre-sorting by hand is needed to separate these from mixed batches of waste batteries, prior to entering the Optisort machine.

The Society for Automotive Engineers and the Battery Association of Japan have both recommended labelling standards for electricvehicle batteries. Recent algorithms from computer vision research have some capability to recognize objects and materials on the basis of features such as size, shape, colour and texture. However, it could be advantageous for recycling if manufacturers were to (some manufacturers already do) include labels, QR Codes, RfID tags or other machinereadable features on key battery components and sub-structures. Where these provide a reference to an external data source, its utility in aiding the recycling process will depend on the accessibility and format of that data. If proprietary and private, such data are of limited use, but there may be initiatives to move towards standardization and open data formats. A number of companies are considering blockchain



**Fig. 2** | Examples of three different battery packs and modules (cylindrical, prismatic and pouch cells) in use in current electric cars. The three designs examined are from model year 2014; this is based on the availability of information from vehicle teardowns, and also because older vehicles are more likely to be closer to end-of-life than today's new cars. The breakdowns include material content in a cell, layout and content of the module and pack and the proportion of critical elements (high economic importance, but at risk of short supply) and strategic materials (either high economic importance or risk of short supply) used. The Nissan pouch cells from Automotive Energy Supply Corporation (AESC) exhibit a mixed cathode chemistry with substantial manganese content and relatively low levels of cobalt. The Tesla cylindrical 18650 cells from Panasonic and the BMW prismatic cells from Samsung SDI both

contain high cobalt levels. Each cell has particular recycling challenges. Cylindrical cells are often bonded into a module using epoxy resin (difficult to remove or recycle); fuses at each end may be blown, making cell discharge challenging; and the cell geometry can be difficult to dismantle for direct recycling. Prismatic cells require 'can opening' (requiring special tools) to remove the contents. These large cells are under considerably more pressure than are the pouch or cylindrical cells, and can therefore be hazardous to open if the contents have degassed. The high manganese content of the Nissan pouch cells makes pyrometallurgical recycling less cost-effective, because manganese is cheap, but these cells are the least problematic to open and physically separate for direct recycling.

technologies to provide whole-life-cycle tracking of battery materials, including information and transparency on provenance, ethical supply chains, battery health and previous use  $^{40}$ . China has signalled its intention to track battery materials.

Automated disassembly of electrical goods has also been implemented to some extent in other sectors. For example, Apple has implemented an automated disassembly line for the iPhone 6<sup>41</sup> that can handle 1.2 million phones per year. This line has 22 stations linked on a conveyor system and can take the iPhone apart in 11 seconds. However, this system can only deal with an iPhone 6 model. Intact phones, of this exact model, must be positioned at the start of the disassembly line, which then uses pre-programmed motions of 29 robots in 21 different cells to dismantle the phone into 8 discrete parts. The LIB is removed by heating the glue which holds the battery in place. Owing to the potential fire hazard,

this must take place inside a thermal event protection system, while monitoring the battery using a thermal imaging system.

Unfortunately, 1.2 million phones per year is a drop in the ocean and the Apple disassembly line has been created using conventional industrial automation methods, making it inflexible and incapable of keeping up adaptively with new models and varieties of phones. But building a flexible and adaptable robot disassembly line need not be prohibitively expensive. The challenge is to create control algorithms and software that can make cheap hardware (robot arms cost only several thousands to several tens of thousands of dollars and costs have been steadily decreasing, can work all the time, and have very long service lifetimes) behave flexibly and intelligently to handle hugely complex disassembly problems. If those artificial intelligence challenges can be solved, then the capital investment required to respond to new and

LIB cathode chemistries		Ideal	• • •	•	• •	•	Poor
Cathode types	LCO	LFP	LMO	NCA	NMC		
Chemical formula	LiCoO <sub>2</sub>	LiFePO <sub>4</sub>	LiMn <sub>2</sub> O <sub>4</sub>	Li(Ni,Co,Al)O <sub>2</sub>	$\begin{aligned} & \text{LiNi}_{0.33}\text{Mn}_{0.33}\text{Co}_{0.33}\text{O}_2 \\ & \text{LiNi}_{0.5}\text{Mn}_{0.3}\text{Co}_{0.2}\text{O}_2 \\ & \text{LiNi}_{0.6}\text{Mn}_{0.2}\text{Co}_{0.2}\text{O}_2 \\ & \text{LiNi}_{0.8}\text{Mn}_{0.1}\text{Co}_{0.1}\text{O}_2 \end{aligned}$	(NMC111) (NMC532) (NMC622) (NMC811)	
Structure	Layered	Olivine	Spinel	Layered	Layered		
Year introduced	1991	1996	1996	1999	2008		
Safety	• •	• • • • •	• • • •	• • •	• • •		
Energy density	• • • •	• • •	• • •	• • • • •	• • • •		
Power density	• • •	• • • •	• • • •	• • • •	• • •		
Calendar lifespan	• • •	• • • •	• • • •	• • • •	• • • •		
Cycle lifespan	• • •	• • • •	• • •	• • • •	• • •		
Performance	• • • •	• • • •	• • •	• • • •	• • • •		
Cost	•	• • • •	• • • •	• • •	• • •		
Market share	Obsolete	Electric bikes, buses and large vehicles	s Small	Steady	Growing (from NMC 1 <sup>-</sup> NMC 811 to no-cobalt		IMC 622 >

Fig. 3 | LIB cathode chemistries. The term LIB covers a range of different battery chemistries, each with different performance attributes. The basic concept of a LIB is that lithium can intercalate into and out of an open structure. which consists of either 'layers' or 'tunnels'. Generally the anode is graphite but the cathode material may have different chemistries and structures, which

result in different performance attributes and there are trade-offs and compromises with each technology. The cathode chemistries of LIBs have a large impact on the performance of LIBs, and these chemistries have evolved and improved. Fig. 3 presents a summary of the different LiB cathode chemistries.

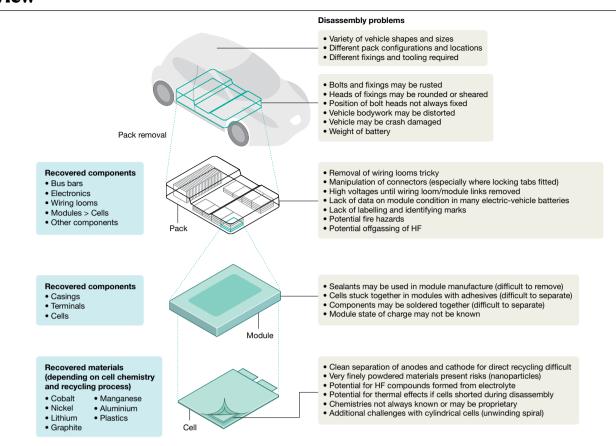
changing models could be kept remarkably low (mainly software updates would be needed). Making robots behave intelligently will rely heavily on sensors to enable advanced robotic perception, especially computer vision using three-dimensional RGB-D imaging devices, combined with bespoke sensors from materials and battery experts. The robots will also require tactile and force-sensing capabilities to handle the complex dynamics problems of forceful interactions between the robots and the materials being disassembled.

Owing to the complexity of automotive battery packs, the possibility of collaborative human-robot co-working using a new generation of force-sensitive 'co-bot' robot arms<sup>33,42</sup> has been suggested. Unlike conventional industrial robots, these co-bots can safely share a workspace with humans, and Wegener<sup>33</sup> suggests that the robot could be taught tasks such as unscrewing bolts, while the human handles cognitively more complex tasks. However, this approach does not protect the human worker from battery hazards and even the task of locating a bolt, moving a tool to engage with it, unscrewing and removing it represents a cutting-edge research challenge in robotics and machine vision. Using current industrial robotics methods, the problem only becomes attemptable (but still difficult) provided that the position of the bolt head is always exactly fixed, in a known pose relative to the robot, with very high precision.

State-of-the-art robotics, computer vision and artificial-intelligence capabilities for handling diverse waste materials do exist, and these systems have demonstrated sufficient robustness and reliability to gain acceptance by the UK nuclear industry, for example, in the deployment of artificial-intelligence-controlled, machine-vision-guided robotic manipulation for cutting of contaminated waste material in radioactive environments<sup>43</sup>. These technologies are now being adapted to the demanding problem of robotic battery disassembly. At different scales of disassembly-pack removal, pack disassembly, module removal and cell separation-different challenges and barriers to automation exist. Some of these are set out in Fig. 4. Computer-vision algorithms are being developed that can identify diverse waste materials and objects<sup>44</sup>, reliably track objects in complex, cluttered scenes<sup>45</sup>, and dynamically guide the actions of robot arms<sup>46</sup>. Dismantling requires forceful interaction between robots and objects, engendering complex dynamics and control problems, such as simultaneous force and motion control<sup>47</sup>, which is needed for robotic cutting or unscrewing. Dismantled materials must be grasped and manipulated, including fragmented or deformable materials, which pose challenges both to vision systems and autonomous grasp planners. Adjigble et al. 48 have recently demonstrated state-of-the-art performance in autonomous, vision-guided robotic grasping of arbitrary objects from random, cluttered heaps. These advances in computer vision, artificial intelligence and robotics fundamentals offer exceptionally promising tools with which to approach the extremely difficult open research challenge of automated disassembly of electric-vehicle batteries.

#### Stabilization and passivation of end-of-life batteries

Once LIBs have been designated for recycling, the three main processes involved consist of stabilization, opening and separation, which may be carried out separately or together. Stabilization of the LIB can be achieved through brine or Ohmic discharge. In-process stabilization during opening, however, is the current route preferred in industry, as it minimizes costs. This consists of shredding or crushing the batteries in an inert gas such as nitrogen, carbon dioxide, or a mixture of carbon dioxide and argon. State-of-the art physical processing of LIBs in Europe and North America includes the Recupyl<sup>8</sup> (France), Akkuser<sup>49</sup> (Finland), Duesenfeld<sup>50</sup> (Germany) and Retriev<sup>51</sup> (USA/Canada) processes. Largescale European processes do not currently use stabilization techniques prior to breaking cells open, instead opting for opening under an inert atmosphere of carbon dioxide or argon (with less than 4% molecular oxygen). Opening under carbon dioxide allows for the formation of a passivating layer of lithium carbonate on any exposed lithium metal. The Retriev process differs from the European processes in that it uses



 $\textbf{Fig. 4} | \textbf{Diagram showing challenges of disassembly at different levels of scale.} \\ \textbf{Electric-vehicle battery packs are complex in design, containing wiring looms, bus bars, electronics, modules, cells and other components.} \\ \textbf{There are } \\ \textbf{There are } \\ \textbf{There } \\ \textbf{T$ 

also many different types of fixtures and fastenings, including screws, bolts, adhesives, sealants and solders, which are not designed for robotic removal.

a water spray during the opening step<sup>51</sup>. The water hydrolyses any exposed lithium and acts as a heatsink, preventing thermal runaway during opening.

Discharging through salt solutions or 'brine' (seawater has been used previously<sup>52,53</sup>) is an alternative method that is supposed to render the cells safe via the corrosion and subsequent water leaching into the cells that passivates the internal cell chemistries. Aqueous solutions of halide salts have been shown to result in substantial corrosion at the battery terminal ends, whereas alkali metal salts, such as sodium phosphate. produce much less corrosion with no water penetration, offering the possibility that cells could be assessed and re-used<sup>53</sup>. This represents a considerably safer discharging method than using seawater; however, competing electrochemical reactions do occur. Oxygen, hydrogen and other gases, such as chlorine (depending upon the salts in the brine), will form at the anode and cathode terminals, and can potentially be collected, though the dangers and difficulties associated with this should not be underestimated. The time for complete discharge is dependent on the solubility of the salt and hence the conductivity of the solution; increasing the temperature will also shorten the discharge time. Once discharge is complete, the cell components can be separated into different materials streams for further processing: steel can or laminated aluminium, separator, anode (graphite, copper, conductive additive), binder and cathode (active material, aluminium, carbon black, binder).

The brine discharge method is not suitable for high-voltage modules and packs, owing to the high rate of electrolysis and vigorous evolution of gases that would occur. However, for low-voltage modules and cells (or once a high-voltage pack has been dismantled into its constituent components) where the electrolysis can be more carefully controlled, this could, in principle, offer a method of discharge in which the hydrogen and oxygen could be recovered for other applications, adding to

the cost-effectiveness of the process<sup>54</sup>. The downside, however, is that contamination of the cell contents threatens to complicate the downstream chemical processes or compromise the value of processed materials streams.

An alternative to the use of salt solutions is direct Ohmic discharge of the battery through a load-bearing circuit. If the electricity can be reclaimed from the discharge, this could offset some of the cost of further processing. To put it into context, the domestic consumption of a standard UK home is up to 4,600 kWh per year. So a 60-kWh battery pack at a 50% state of charge and a 75% state of health has a potential 22.5 kWh for end-of-life reclamation, which would power a UK home for nearly 2 hours. At 14.3 p per kWh, this equates to UK£3.22 per pack, which may seem a modest gain that does not warrant the cost of investing in equipment. However, if it is unrecovered, the energy from discharge must be dissipated, and this will add to the cooling burden of the facility, creating additional costs. Furthermore, an economy of scale is to be anticipated when recycling electric vehicle batteries in bulk. Similarly, reclaimed energy might make a useful contribution to the profitability of repurposing for second use (see section 'Battery assessment and disassembly').

LIB cells can be shredded at various states of charge, and from a commercial point of view, if discharged modules or cells are to be processed in this way, discharge prior to shredding adds cost to the processes. Furthermore, exactly what the optimum level of discharge might be remains unclear. Depending on cell chemistry and depth of discharge, over-discharging of cells can result in copper dissolution into the electrolyte. The presence of this copper is detrimental for materials reclamation as it may then contaminate all the different materials streams, including the cathode and separator. If the voltage is then increased again or 'normal' operation resumed<sup>55</sup>, this can be dangerous because

copper can reprecipitate throughout the cell, increasing the risks of short-circuiting and thermal runaway.

Current LIB-processing technologies essentially bypass these concerns by feeding end-of-life batteries directly into a shredder or high-temperature reactor. Industrial comminution technologies can passivate batteries directly but recovered battery materials then require a complex set of physical and chemical processes to produce usable materials streams. Pyrometallurgical recycling processes (see section 'Stabilization and passivation of end-of-life batteries') at scale may be able to accept entire electric-vehicle modules without further disassembly. However, this solution fails to capture much of the embodied energy that goes into LIB manufacture, and leaves chemical separation techniques with much to do as the battery materials become ever more intimately mixed.

#### **Recycling methods**

#### Pyrometallurgical recovery

Pyrometallurgical metals reclamation uses a high-temperature furnace to reduce the component metal oxides to an alloy of Co, Cu, Fe and Ni. The high temperatures involved mean that the batteries are 'smelted', and the process, which is a natural progression from those used for other types of batteries, is already established commercially for consumer LIBs. It is particularly advantageous for the recycling of general consumer LIBs, which currently tends to be geared towards an imperfectly sorted feedstock of cells (indeed, the batteries can be processed along with other types of waste to improve the thermodynamics and products obtained), and this versatility is also valuable with respect to electricvehicle LIBs. As the metal current collectors aid the smelting process<sup>56</sup>. the technique has the important advantage that it can be used with whole cells or modules, without the need for a prior passivation step.

The products of the pyrometallurgical process are a metallic alloy fraction, slag and gases. The gaseous products produced at lower temperatures (<150 °C) comprise volatile organics from the electrolyte and binder components. At higher temperatures the polymers decompose and burn off. The metal alloy can be separated through hydrometallurgical processes (see section 'Hydrometallurgical metals reclamation') into the component metals, and the slag typically contains the metals aluminium, manganese and lithium, which can be reclaimed by further hydrometallurgical processing, but can alternatively be used in other industries such as the cement industry. There is relatively little safety risk in this process, as the cells and modules are all taken to extreme temperatures with a reductant for metal reclamation—aluminium from the electrode foils and packaging is a major contributor here—so the hazards are contained within the processing. In addition, the burning of the electrolytes and plastics is exothermic and reduce the energy consumption required for the process. It follows that in the pyrometallurgical process there is typically no consideration given to the reclamation of the electrolytes and the plastics (approximately 40–50 per cent of the battery weight) or other components such as the lithium salts. Despite environmental drawbacks (such as the production of toxic gases, which must be captured or remediated and the requirement for hydrometallurgical post-processing), high energy costs, and the limited number of materials reclaimed, this remains a frequently used process for the extraction of high-value transition metals such as cobalt and nickel<sup>57</sup>.

#### Physical materials separation

For reclamation after comminution, recovered materials can be subjected to a range of physical separation processes that exploit variations in properties such as particle size, density, ferromagnetism and hydrophobicity. These processes include sieves, filters, magnets, shaker tables and heavy media, used to separate a mixture of lithium-rich solution, lowdensity plastics and papers, magnetic casings, coated electrodes and electrode powders. The result is generally a concentration of electrode coatings in the fine fractions of material, and a concentration of plastics,

casing materials, and metal foils in the coarse fractions<sup>58</sup>. The coarse fractions can be put through magnetic separation processes to remove magnetic material such as steel casings and density separation processes to separate plastics from foils. The fine product is referred to as the 'black mass', and comprises the electrode coatings (metal oxides and carbon). The carbon can be separated from metal oxides by froth flotation, which exploits the hydrophobicity of carbon to separate it from the more hydrophilic metal oxides<sup>59</sup>. An overview of how these processes are used by several companies is shown in Fig. 5, which mentions the Recupyl<sup>8</sup> (France), Akkuser<sup>49</sup> (Finland), Duesenfeld<sup>50</sup> (Germany) and Retriev<sup>51</sup> (USA/ Canada) processes.

Often, the polymeric binders from the 'black mass' components need to be eliminated to liberate the graphite and metal oxides from the copper and aluminium current collectors. Published routes include the use of sonication in a solvent such as N-methyl-2-pyrrolidone (NMP) or dimethylformamide (DMF) to detach the cathode from the current collector<sup>60</sup>, thermal heat treatment to decompose the binder<sup>61,62</sup>, or dis $solution of the aluminium current collector {}^{63}. These processes, however, \\$ often require high temperatures (60-100 °C) and are relatively slow (3 h). While ultrasound can induce faster delamination (1.5 h), this is still too slow for a continuous-flow process and the required solventto-solid mass ratios of 10:1 will not be viable on a commercial scale with these solvents<sup>64</sup>.

Recent teardowns of cells indicate that manufacturers are transitioning away from fluorinated binders. Many newer batteries are moving toward alternative binders on the anode, such as carboxymethyl cellulose (CMC), which is water-soluble, and styrene butadiene rubber (SBR), which is not water-soluble but is applied as an emulsion that may be easier to remove at end-of-life. There is also work on water-based binder systems for cathodes, but this is proving to be more challenging. Other studies have used cellulose- and lignin-based binders, although many of these are still in the laboratory testing phase<sup>65</sup>.

#### Hydrometallurgical metals reclamation

Hydrometallurgical treatments involve the use of aqueous solutions to leach the desired metals from cathode material. By far the most common combination of reagents reported is H<sub>2</sub>SO<sub>4</sub>/H<sub>2</sub>O<sub>2</sub> (ref. <sup>66</sup>). A number of studies have been carried out in order to determine the most efficient set of conditions to achieve an optimal leaching rate. These include: concentration of leaching acid, time, temperature of solution, the solidto-liquid ratio and the addition of a reducing agent <sup>67</sup>. In most of these studies, it was found that leaching efficiency improved when H<sub>2</sub>O<sub>2</sub> was added. Somewhat counterintuitively, it is understood that H<sub>2</sub>O<sub>2</sub> acts as a reducing agent to convert insoluble Co(III) materials into soluble Co(II) through the reaction<sup>7</sup>:

$$2LiCoO_2(s)+3H_2SO_4+H_2O_2 \rightarrow 2CoSO_4(aq)+Li_2SO_4+4H_2O+O_2$$

A range of other possible leaching acids and reducing agents have been investigated<sup>68-72</sup>. The leached solution may also subsequently be treated with an organic solvent to perform a solvent extraction<sup>73-75</sup>. Once leached, the metals may be recovered through a number of precipitation reactions controlled by manipulating the pH of the solution. Cobalt is usually extracted either as the sulfate, oxalate, hydroxide or  $carbonate^{75-79}, and \, then \, lithium \, can \, be \, extracted \, through \, a \, precipitation$ reaction forming Li<sub>2</sub>CO<sub>3</sub> or Li<sub>3</sub>PO<sub>4</sub><sup>80,81</sup>. An alternative recycling method describes mechanochemical treatment of materials, where electrode materials are ground with a chlorine compound or complexing agent to produce water-soluble salts of cobalt, which can be separated from insoluble fractions by washing with water<sup>82,83</sup>.

Most current recycling processes fall under the umbrella of 'reagent recovery' because the materials, with sufficient purity, can be re-used not just for resynthesizing the original cathode materials, but also in a range of other applications, such as the synthesis of CoFe<sub>2</sub>O<sub>4</sub> or MnCo<sub>2</sub>O<sub>4</sub> (refs. 84-86). Following initial work focused on the leaching and

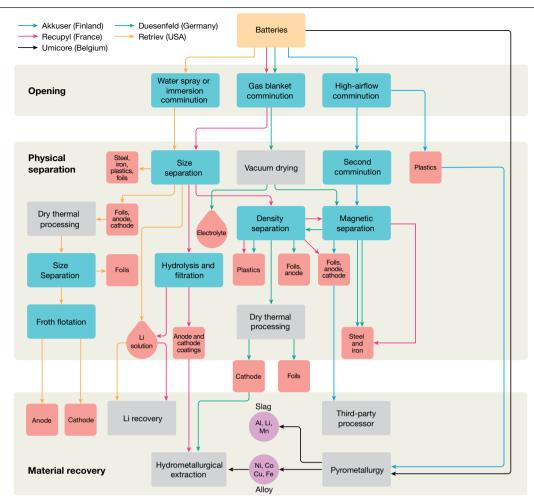


Fig. 5 | Flow chart representing potential routes for the circular economy of LIBs, detailing second-use applications, re-use, physical recovery, chemical recovery and biorecovery. A range of commercial entities have

commercialized processes for recycling LIBs. Different approaches for the physical separation of batteries and the recovery of materials are indicated.

remanufacture of LiCoO<sub>2</sub> (ref. 87), work has since moved on to strategies for new cell chemistries, which typically contain multiple transition metals (for example, LiNi<sub>1-x-y</sub>Mn<sub>x</sub>Co<sub>y</sub>O<sub>2</sub>; NMC). In such cases, once the metals have been leached from the cathode material, either sequential precipitation is employed to recover the individual metals, or the direct remanufacture of the cathode is targeted, such as work to recover NMC88. In this work, after leaching the metals from the cathode, the concentrations of the various metals in solution were measured and adjusted to match those in the target material (1:1:1 Ni:Mn:Co for NMC-111). The same group has applied the technique to NMC with varying metal contents and successfully resynthesized such NMC materials through the production of a precursor hydroxide, Ni<sub>x</sub>Mn<sub>y</sub>Co<sub>z</sub>(OH)<sub>2</sub> with x, y and z varying according to the desired final composition of the cathode89.

Other groups have published similar recovery methods with modifications such as additional solvent extraction steps90, lactic acid or urea as an alternative to sulfuric acid (additionally facilitating resynthesis)<sup>91,92</sup> as well as investigating the effect of magnesium in the resynthesized material93. The big issues to be addressed with all solvo-metallurgical processes are the volumes of solvents required, the speed of delamination, the costs of neutralization and the likelihood of cross-contamination of materials. Although shredding is a fast and efficient method of rendering the battery materials safe, mixing the anode and cathode materials at the start of the recycling process complicates downstream processing. A method in which anode and cathode assemblies could be separated prior to mechanical or solvent-based separation would greatly improve material segregation. This is one of several key areas where designing for end-of-life recycling promises to have a real impact, but the historic backlog of batteries containing polyvinylidene fluoride (PVDF) as a binder will still need to be processed. It is clear that the current design of cells makes recycling extremely complex and neither hydro-nor pyrometallurgy currently provides routes that lead to pure streams of material that can easily be fed into a closed-loop system for batteries.

#### **Direct recycling**

The removal of cathode or anode material from the electrode for reconditioning and re-use in a remanufactured LIB is known as direct recycling. In principle, mixed metal-oxide cathode materials can be reincorporated into a new cathode electrode with minimal changes to the crystal morphology of the active material. In general, this will require the lithium content to be replenished to compensate for losses due to degradation of the material during battery use and because materials may not be recovered from batteries in the fully discharged state with the cathodes fully lithiated. So far, work in this area has focused primarily on laptop and mobile phone batteries, as a result of the larger amounts of these available for recycling<sup>38</sup>. An example of how this recycling route could work has been outlined recently94. Cathode strips, obtained after dismantling spent batteries, were soaked in NMP before undergoing sonication. Powders were either regenerated through simple solid-state synthesis with the addition of fresh Li<sub>2</sub>CO<sub>3</sub> or treated hydrothermally with a solution containing LiOH/Li<sub>2</sub>SO<sub>4</sub> before annealing.

For high-cobalt cathodes such as lithium cobalt oxide (LCO) conventional pyrometallurgical (see section 'Pyrometallurgical recovery') or

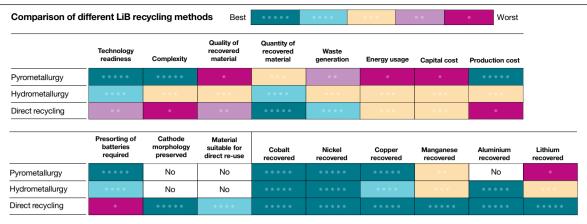


Fig. 6 | Comparison of different LiB recycling methods.

hydrometallurgical (see section 'Hydrometallurgical recovery') recycling processes can recover around 70% of the cathode value<sup>11</sup>. However, for other cathode chemistries that are not as cobalt-rich, this figure drops notably<sup>11</sup>. A 2019 648-lb Nissan Leaf battery, for example, costs US\$6,500-8,500 new, but the value of the pure metals in the cathode material is less than US\$400 and the cost of the equivalent amount of NMC (an alternative cathode material) is in the region of US\$4,000. It is important, therefore, to appreciate that cathode material must be directly recycled (or upcycled) to recover sufficient value. As direct recycling avoids lengthy and expensive purification steps, it could be particularly advantageous for lower-value cathodes such as LiMn<sub>2</sub>O<sub>4</sub> and LiFePO<sub>4</sub>, where manufacturing of the cathode oxides is the major contributor to cathode costs, embedded energy and carbon dioxide footprint95.

Direct recycling also has the advantage that, in principle, all battery components<sup>20</sup> can be recovered and re-used after further processing (with the exclusion of separators). Although there is substantial literature regarding the recycling of the cathode component from spent LIBs, research on recycling of the graphitic anode is limited, owing to its lower recovery value. Nevertheless, the successful re-use of mechanically separated graphite anodes from spent batteries has been demonstrated, with similar properties to that of pristine graphite 96.

Despite the potential advantages of direct recycling, however, considerable obstacles remain to be overcome before it can become a practical reality. The efficiency of direct recycling processes is correlated with the state of health of the battery and may not be advantageous where the state of charge is low<sup>97</sup>. There are also potential issues with the flexibility of these routes to handle metal oxides of different compositions. For maximum efficiency, direct recycling processes must be tailored to specific cathode formulations, necessitating different processes for different cathode materials<sup>97</sup>. The ten or so years spent in a vehicle-followed, perhaps, by a few more in a second-use application—therefore present a challenge in an industry where battery formulations are evolving at a rapid pace. Direct recycling may struggle to accommodate feedstocks of unknown or poorly characterized provenance, and there will be commercial reluctance to re-use material if product quality is affected.

The direct recycling route for cathode coatings is also highly sensitive to contamination by other metals, such as aluminium, which results in poor electrochemical performance<sup>60</sup>. In particular, methods of recovering materials for further physical or chemical separation that involve a high degree of comminution form fine particles of Al and Cu, which are difficult to separate from the electrode coatings. For this reason, processes that do not mechanically stress the electrode foils are favoured in direct recycling, and separation of the materials streams prior to mechanical sorting is preferable. However, methods of removing the electrode binder-typically pyrolysis or dissolution-present further

challenges, such as the production of hazardous byproducts such as HF from pyrolysis of the PVDF binder or the use of the highly toxic NMP as a solvent for dissolution. The potential for the undesirable reaction of the PVDF binder with the electrode material appears to be a notable omission in the recycling literature, despite a growing body of research illustrating that PVDF is an excellent low-temperature fluorinating reagent for metal oxides 98. Furthermore, recent research suggests that a certain degree of reaction can occur with the cathode even under conditions of normal cell operation<sup>99</sup>.

## **Biological metals reclamation**

Bioleaching, in which bacteria are harnessed to recover valuable metals, has been used successfully in the mining industry 100,101. This is an emerging technology for LIB recycling and metal reclamation and is potentially complementary to the hydrometallurgical and pyrometallurgical processes currently used for metal extraction 102,103; cobalt and nickel, in particular, are difficult to separate and require additional solvent-extraction steps. The process uses microorganisms to digest metal oxides from the cathode selectively 104 and to reduce these oxides to produce metal nanoparticles 105,106. The number of studies that have been performed thus far, however, is relatively small and there is plenty of opportunity for further investigation in this field. The recycling methods discussed are compared in Fig. 6.

# **Summary and opportunities**

The electric-vehicle revolution is set to change the automotive industry radically, and some of the most profound changes will inevitably relate to the management and decommissioning of vehicles at end-of-life. Of chief concern are the complex, high-tech power trains and, in particular, the LIBs. To put this into perspective, electrification of only 2% of the current global car fleet would represent a line of cars-and in due course, of end-of-life waste-that could stretch around the Earth. There is wide acceptance that, for environmental and safety reasons, stockpiling (or worse, landfill) and wholesale transport of end-of-life electric-vehicle batteries are not attractive options, and that the management of endof-life electric-vehicle waste will require regional solutions.

In the waste management hierarchy, re-use is considered preferable to recycling, in order to extract maximum economic value and minimize environmental impacts. Many companies in various parts of the world are already piloting the second use of electric-vehicle LIBs for a range of energy storage applications. Advanced sensors and improved methods of monitoring batteries in the field and end-of-life testing would enable the characteristics of individual end-of-life batteries to be better matched to proposed second-use applications, with concomitant advantages in lifetime, safety and market value. Even if all the benefits of

# Review

second-use are realized, however, it must be remembered that recycling (if not landfill) is the inevitable fate of all batteries.

Some recent life-cycle analyses has indicated that the application of current recycling processes to the present generation of electric-vehicle LIBs may not in all cases result in reductions in greenhouse gas emissions compared to primary production<sup>107</sup>. More efficient processes are urgently needed to improve both the environmental and economic viability of recycling, which at present is heavily dependent on cobalt content. However, as the amount of cobalt in cathodes is reduced for economic and other reasons, to recycle using current methods will become less advantageous owing to the lower value of the materials recovered.

At present, there are low volumes of electric-vehicle batteries that require recycling. As these volumes increase dramatically, there are questions concerning the economies (and diseconomies) of scale in relation to recycling operations<sup>58</sup>. Pyrometallurgical routes, in particular, suffer from high capital costs, and if full recyclability of LIBs is to be achieved, alternative methods are urgently required, rather than seeking to recycle only the most economically valuable components.

There are a number of lessons that the future LIB recycling industry could learn from the highly successful lead-acid battery recycling industry. As a technology, lead-acid batteries are relatively standardized and simple to disassemble and recycle, which minimizes costs, allowing the value of lead to drive recycling. Unfortunately, for a rapidly developing technology such as electric-vehicle LIBs, such advantages are not likely to apply any time soon.

A number of improvements could make electric-vehicle LIB recycling processes economically more efficient<sup>23</sup>, such as better sorting technologies, a method for separating electrode materials, greater process flexibility, design for recycling, and greater manufacturer standardization of batteries. There is a clear opportunity for a more sophisticated approach to battery recovery through automated disassembly, smart segregation of different batteries and the intelligent characterization, evaluation and 'triage' of used batteries into streams for remanufacture, re-use and recycling. The potential benefits of this are many and include reduced costs, higher value of recovered material streams, and the near elimination of the risk of harm to human workers.

The design of current battery packs is not optimized for easy disassembly. Use of adhesives, bonding methods and fixtures do not lend themselves to easy deconstruction either by hand or machine. All reported current commercial physical cell-breaking processes employ shredding or milling with subsequent sorting of the component materials. This makes the separation of the components more difficult than if they were presorted and considerably reduces the economic value of waste material streams. Many of the challenges this presents to remanufacture, re-use and recycling could be addressed if considered early in the design process.

For direct recycling where purity of the recovered materials is required, a process which involves less component contamination during the breaking stage is important. This would benefit from an analysis of the cell component chemistries, and the state of charge and state of health of the cells before disassembly into the component parts, rather than the production of a mixture of all components. At present, this separation has only been performed at a laboratory scale and usually employs manual disassembly methods that are difficult to scale up economically. The move to greater automation and robotic disassembly promises to overcome some of these hurdles. Issues regarding the binder still need to be resolved, and acid, alkali, solvent and thermal treatments all have their positives and negatives. A cell design for reclamation of materials is extremely appealing, with low-cost water-soluble binders.

We have focused here on the scientific challenges of recycling LIBs, but we recognize that the 'system performance' of the LIB recycling industry will be strongly affected by a range of non-technical factors, such as the nature of the collection, transportation, storage and logistics of LIBs at the end-of-life. As these vary from country to country and region to region, it follows that different jurisdictions may arrive

at different answers to the problems posed. Research is under way in the Faraday Institution ReLiB Project, UK: the ReCell Project, US: at CSIRO in Australia and at a number of European Union projects including ReLieVe, Lithorec and AmplifII.

Recycling electric-vehicle batteries at end-of-life is essential for many reasons. At present there is little hope that profitable processes will be found for all types of current and future types of electric-vehicle LIBs without substantial successful research and development, so the imperative to recycle will derive primarily from the desire to avoid landfill and to secure the supply of strategic elements. The environmental and economic advantages of second-use and the low volume of electric-vehicle batteries currently available for recycling could stifle the development of a recycling industry in some places. In many nations, the elements and materials contained in the batteries are not available, and access to resources is crucial in ensuring a stable supply chain. Electric vehicles may prove to be a valuable secondary resource for critical materials. Careful husbandry of the resources consumed by electric-vehicle battery manufacturing-and recycling-surely hold the key to the sustainability of the future automotive industry.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1682-5.

- International Energy Agency (IEA) Global EV Outlook 2018 (IEA, 2018).
- Ahmadi, L., Young, S. B., Fowler, M., Fraser, R. A. & Achachlouei, M. A. A cascaded life cycle: reuse of electric vehicle lithium-ion battery packs in energy storage systems. Int. J. Life Cycle Assess. 22, 111-124 (2017).
- Doughty, D. H. & Roth, E. P. A general discussion of Li ion battery safety. Electrochem. Soc. Interface 21, 37-44 (2012).
- Kong, L., Li, C., Jiang, J. & Pecht, M. Li-ion battery fire hazards and safety strategies. Energies 11, 2191 (2018).
- Rethink Waste https://www.rethinkwaste.org/uploads/media\_items/111617-shorewayoperations.original.pdf (Shoreway Operations and Contract Management, 2017).
- Reaugh, L. American Manganese: Virtual Reality International Conference (VRIC) Conversation with President and CEO Larry Reaugh - MoonShot Exec, https:// moonshotexec.com/american-manganese-vric-conversation-with-president-and-ceolarry-reaugh/ (2018).
- Meshram, P., Pandey, B. D. & Mankhand, T. R. Extraction of lithium from primary and secondary sources by pre-treatment, leaching and separation: a comprehensive review. Hydrometallurgy 150, 192-208 (2014).
- Tedjar, F. in Challenge for Recycling Advanced EV Batteries https://congresses.icmab.es/ iba2013/images/files/Friday/Morning/Farouk%20Tedjar.pdf (2013).
- Katwala, A. The spiralling environmental cost of our lithium battery addiction. Wired https://www.wired.co.uk/article/lithium-batteries-environment-impact (2018).
- Larcher, D. & Tarascon, J.-M. Towards greener and more sustainable batteries for electrical energy storage, Nat. Chem. 7, 19-29 (2015).
- Gaines, L. Lithium-ion battery recycling processes: research towards a sustainable course, Sustain, Mater, Technol, 17, e00068 (2018)
- The net impact of LIB production can be greatly reduced if more materials can be recovered from end-of-life LIBs, in as usable a form as possible.
- Turcheniuk, K., Bondarev, D., Singhal, V. & Yushin, G. Ten years left to redesign lithium-ion batteries, Nature 559, 467-470 (2018),
- Tahil, W. The Trouble with Lithium: Implications of Future PHEV Production for Lithium Demand (Meridian International Research, 2007).
- Gaines, L. & Nelson, P. Lithium-ion batteries: examining material demand and recycling issues. In TMS 2010 Annual Meeting and Exhibition 27-39 (TMS 2013). Initial concerns regarding resource constraints for scaling up LIB production focused on lithium; however, in the near term, reserves of lithium are unlikely to present a constraint
- Narins, T. P. The battery business: lithium availability and the growth of the global electric car industry. Extr. Ind. Soc. 4, 321-328 (2017).
- Schmuch, R., Wagner, R., Hörpel, G., Placke, T. & Winter, M. Performance and cost of materials for lithium-based rechargeable automotive batteries. Nat. Energy 3, 267 (2018)
- Nkulu, C. B. L. et al. Sustainability of artisanal mining of cobalt in DR Congo. Nat. Sustain. 1, 495 (2018).
- Gür, T. M. Review of electrical energy storage technologies, materials and systems: challenges and prospects for large-scale grid storage. Energy Environ. Sci. 11, 2696-2767 (2018)
- Sun, S. I., Chipperfield, A. J., Kiaee, M. & Wills, R. G. A. Effects of market dynamics on the time-evolving price of second-life electric vehicle batteries. J. Energy Storage 19, 41-51

- Gaines, L. The future of automotive lithium-ion battery recycling: charting a sustainable course. Sustain. Mater. Technol. 1-2, 2-7 (2014).
- Jaffe, S. Vulnerable links in the lithium-ion battery supply chain. Joule 1, 225-228 (2017).
- 22. Helbig, C., Bradshaw, A. M., Wietschel, L., Thorenz, A. & Tuma, A. Supply risks associated with lithium-ion battery materials. J. Clean. Prod. 172, 274-286 (2018). Focusing on six battery systems (LCO-C, LMO-C, NMC-C, NCA-C, LFP-C and LFP-LTO)
  - this research evaluates the relative supply risk for individual elements (Li, Al, Ti, Mn, Fe, Co. Ni, Cu. P and graphite) in LIBs.
- Diekmann, J. et al. Ecological recycling of lithium-ion batteries from electric vehicles with focus on mechanical processes. J. Electrochem. Soc. 164, A6184-A6191 (2017).
- Nedjalkov, A. et al. Toxic gas emissions from damaged lithium ion batteries—analysis and safety enhancement solution. Batteries 2, 5 (2016).
- Elwert, T., Römer, F., Schneider, K., Hua, Q. & Buchert, M. in Behaviour of Lithium-Ion Batteries in Electric Vehicles (eds Pistoia, G. & Liaw, B.) 289-321 (Springer, 2018). This article describes the recycling and value chain of LIBs from vehicles and the different industrial approaches currently used for cell recycling, discussing the economic and ecological aspects briefly and highlighting current challenges of LIB recycling.
- Lambert, S. M. et al. Rapid nondestructive-testing technique for in-line quality control of Li-ion batteries, IEEE Trans, Ind. Electron, 64, 4017-4026 (2017).
- 27 Attidekou, P. S., Wang, C., Armstrong, M., Lambert, S. M. & Christensen, P. A. A New Time Constant Approach to Online Capacity Monitoring and Lifetime Prediction of Lithium Ion Batteries for Electric Vehicles (EV). J. Electrochem. Soc. 164, A1792-A1801 (2017).
- Attidekou, P. S. et al. A study of 40 Ah lithium ion batteries at zero percent state of charge as a function of temperature. J. Power Sources 269, 694-703 (2014).
- Cerdas, F. et al. in Recycling of Lithium-Ion Batteries 83-97 (Springer, 2018).
- 30. Institute of the Motor Industry (IMI) IMI Raises Skills And Regulation Concerns As Demand For Electric And Hybrid Vehicle Surges https://www.theimi.org.uk/news/imi-raises-skillsand-regulation-concerns-demand-electric-and-hybrid-vehicle-surges (IMI, 2015)
- EVs and industrial strategy. In Electric Vehicles: Driving The Transition https://publications. parliament.uk/pa/cm201719/cmselect/cmbeis/383/38309.htm. (Business, Energy and Industrial Strategy Committee, House of Commons, UK, 2018).
- 32. Duflou, J. R. et al. Efficiency and feasibility of product disassembly: a case-based study. CIRP Ann. 57, 583-600 (2008).
- Wegener, K., Chen, W. H., Dietrich, F., Dröder, K. & Kara, S. Robot assisted disassembly for the recycling of electric vehicle batteries. Proc. CIRP 29, 716-721 (2015).
- 34. Dornfeld, D. A. & Linke, B. S. (eds) Leveraging Technology for a Sustainable World. (Proc. 19th CIRP Conf. on Life Cycle Engineering) (Springer, 2012).
- Markowski, J., Ay, P., Pempel, H. & Müller, M. in Recycling und Rohstoffe https://www.vivis. 35. de/wp-content/uploads/RuR5/2012\_RuR\_443\_456\_Markowski.pdf (TK, 2012).
- 36. ReLiB. Gateway Testing & Dismantling. https://relib.org.uk/gateway-testing-dismantling/ (The Faraday Institution, 2019).
- 37. Arora, S. & Kapoor, A. in Behaviour of Lithium-Ion Batteries in Electric Vehicles (eds Pistoia. G & Liaw B ) 175-200 (Springer 2018)
- Chen, H. & Shen, J. A degradation-based sorting method for lithium-ion battery reuse. 38. PLoS One 12, e0185922 (2017).
- Advances in Battery Technologies for Electric Vehicles (eds Bruno Scrosati, B., Jürgen 39 Garche, J. & Werner Tillmetz, W.) 245-263 (Elsevier, 2015).
- 40 Bazilian, M. D. The mineral foundation of the energy transition. Extr. Ind. Soc. 5, 93-97 (2018).
- 41 Rujanavech, C. et al. Liam-An Innovation Story (Apple, 2016).
- Luca, A., Albu-Schaffer, A., Haddadin, S. & Hirzinger, G. in 2006 IEEE/RSJ Int. Conf. on 42. Intelligent Robots and Systems 1623-1630 (IEEE, 2006).
- Chapman, H., Lawton, S. & Fitzpatrick, J. Laser cutting for nuclear decommissioning: an integrated safety approach. Atw. Int. Z. Kernenergie 63, 521-526 (2018).
- Sun, L. et al. A novel weakly-supervised approach for RGB-D-based nuclear waste object detection. IEEE Sens. J. 19, 3487-3500 (2018).
- Xiao, J., Stolkin, R., Gao, Y. & Leonardis, A. Robust fusion of color and depth data for RGB-D target tracking using adaptive range-invariant depth models and spatio-temporal consistency constraints. IEEE Trans. Cybern. 48, 2485-2499 (2018).
- 46. Marturi, N. et al. Dynamic grasp and trajectory planning for moving objects. Auton. Robots 43, 1241-1256 (2018).
- Ortenzi, V., Stolkin, R., Kuo, J. & Mistry, M. Hybrid motion/force control: a review. Adv. Robot. 31, 1102-1113 (2017).
- Adjigble, M. et al. Model-free and learning-free grasping by Local Contact Moment 48. matching. In Int. Conf. on Intelligent Robots and Systems (IROS) 2933-2940 (IEEE, 2018). This paper presents an algorithm that is key to automated battery processing, in which an artificial intelligence and robotic vision system can autonomously plan where to place a robot's fingers to stably grasp an arbitrarily shaped object, without relying on any prior knowledge or models of the object or needing any machine learning using offline training data.
- 49. Pudas, J., Erkkila, A. & Viljamaa, J. Battery recycling method. US Patent No. 8, 979, 006 (2010)
- 50. Hanisch, C. Recycling method for treating used batteries, in particular rechargeable batteries, and battery processing installation. US Patent Application 2019/0260101A1 (2019).
- Smith, W. N. & Swoffer, S. Recovery of lithium ion batteries. US Patent 8, 616, 475 (2013).
- Li, J., Wang, G. & Xu, Z. Generation and detection of metal ions and volatile organic compounds (VOCs) emissions from the pretreatment processes for recycling spent lithium-ion batteries. Waste Manag. 52, 221-227 (2016).
- Shaw-Stewart, J. et al. Aqueous solution discharge of cylindrical lithium-ion cells. Sustain. Mater. Technol. https://doi.org/10.1016/j.susmat.2019.e00110 (2019).
- Al-Thyabat, S., Nakamura, T., Shibata, E. & Iizuka, A. Adaptation of minerals processing operations for lithium-ion (LiBs) and nickel metal hydride (NiMH) batteries recycling: critical review. Miner. Eng. 45, 4-17 (2013).
- Guo, R., Lu, L., Ouyang, M. & Feng, X. Mechanism of the entire overdischarge process and overdischarge-induced internal short circuit in lithium-ion batteries. Sci. Rep. 6, 30248 (2016)

- Georgi-Maschler, T., Friedrich, B., Weyhe, R., Heegn, H. & Rutz, M. Development of a recycling process for Li-ion batteries. J. Power Sources 207, 173-182 (2012)
- 57. Lv, W. et al. A critical review and analysis on the recycling of spent lithium-ion batteries. ACS Sustain. Chem. Eng. 6, 1504-1521 (2018).
- Wang, X., Gaustad, G. & Babbitt, C. W. Targeting high value metals in lithium-ion battery recycling via shredding and size-based separation. Waste Manag. 51, 204-213
- Zhan, R., Oldenburg, Z. & Pan, L. Recovery of active cathode materials from lithium-ion batteries using froth flotation. Sustain. Mater. Technol. 17, e00062 (2018).
- 60. Li, X., Zhang, J., Song, D., Song, J. & Zhang, L. Direct regeneration of recycled cathode material mixture from scrapped LiFePO<sub>4</sub> batteries. J. Power Sources 345, 78-84 (2017).
- Li, J., Wang, G. & Xu, Z. Environmentally-friendly oxygen-free roasting/wet magnetic separation technology for in situ recycling cobalt, lithium carbonate and graphite from spent LiCoO<sub>2</sub>/graphite lithium batteries, J. Hazard, Mater. 302, 97-104 (2016).
- Song, D. et al. Recovery and heat treatment of the Li(Ni<sub>1/3</sub>Co<sub>1/3</sub>Mn<sub>1/3</sub>)O<sub>2</sub> cathode scrap material for lithium ion battery. J. Power Sources 232, 348-352 (2013).
- Chen, J. et al. Environmentally friendly recycling and effective repairing of cathode 63. powders from spent LiFePO, batteries, Green Chem. 18, 2500-2506 (2016).
- Zhang, Z, et al. Ultrasound-assisted hydrothermal renovation of LiCoO<sub>2</sub> from the cathode of spent lithium-ion batteries. Int. J. Electrochem. Sci. 9, 3691-3700 (2014).
- 65 Nirmale, T. C., Kale, B. B. & Varma, A. J. A review on cellulose and lignin based binders and electrodes: small steps towards a sustainable lithium ion battery. Int. J. Biol. Macromol. 103 1032-1043 (2017)
- 66 Ferreira, D. A., Prados, L. M. Z., Majuste, D. & Mansur, M. B. Hydrometallurgical separation of aluminium, cobalt, copper and lithium from spent Li-ion batteries. J. Power Sources 187. 238-246 (2009).
- He, L.-P., Sun, S.-Y., Song, X.-F. & Yu, J.-G. Leaching process for recovering valuable metals from the  $LiNi_{1/3}Co_{1/3}Mn_{1/3}O_2$  cathode of lithium-ion batteries. Waste Manag. 64, 171–181 (2017).
- Li, J., Shi, P., Wang, Z., Chen, Y. & Chang, C.-C. A combined recovery process of metals in 68. spent lithium-ion batteries. Chemosphere 77, 1132-1136 (2009).
- Nayaka, G. P., Pai, K. V., Santhosh, G. & Manjanna, J. Dissolution of cathode active material of spent Li-ion batteries using tartaric acid and ascorbic acid mixture to recover Co. Hydrometallurgy 161, 54-57 (2016).
- Pinna, E. G, Ruiz, M. C., Ojeda, W. M. & Rodriguez, M. H. Cathodes of spent Li-ion batteries: dissolution with phosphoric acid and recovery of lithium and cobalt from leach liquors. Hydrometallurgy 167, 66-71 (2016).
- Yang, L. et al. Preparation and magnetic performance of Co., Fe., O. by a sol-gel method using cathode materials of spent Li-ion batteries. Ceram. Int. 42. 1897-1902 (2016)
- Zheng, X, et al. Spent lithium-ion battery recycling—reductive ammonia leaching of metals from cathode scrap by sodium sulphite. Waste Manag. 60, 680-688 (2017).
- Granata, G., Moscardini, E., Pagnanelli, F., Trabucco, F. & Toro, L., Product recovery from Li-ion battery wastes coming from an industrial pre-treatment plant; lab scale tests and process simulations. J. Power Sources 206, 393-401 (2012).
- Mantuano, D. P., Dorella, G., Elias, R. C. A. & Mansur, M. B. Analysis of a hydrometallurgical route to recover base metals from spent rechargeable batteries by liquid-liquid extraction with Cyanex 272. J. Power Sources 159, 1510-1518 (2006).
- Kang, J., Senanayake, G., Sohn, J. & Shin, S. M. Recovery of cobalt sulfate from spent lithium ion batteries by reductive leaching and solvent extraction with Cyanex 272. Hydrometallurgy 100, 168-171 (2010).
- Kang, J., Sohn, J.-S., Chang, H., Senanayake, G. & Shin, S. Preparation of cobalt oxide from concentrated cathode material of spent lithium ion batteries by hydrometallurgical method. Adv. Powder Technol. 21, 175-179 (2010).
- Pagnanelli, F., Moscardini, E., Altimari, P., Abo Atia, T. & Toro, L. Cobalt products from real waste fractions of end of life lithium ion batteries. Waste Manag. 51, 214-221
- Hu, C., Guo, J., Wen, J. & Peng, Y. Preparation and electrochemical performance of nano-Co<sub>3</sub>O<sub>4</sub> anode materials from spent Li-ion batteries for lithium-ion batteries. J. Mater. Sci. Technol. 29, 215-220 (2013).
- Paulino, J. F., Busnardo, N. G. & Afonso, J. C. Recovery of valuable elements from spent Li-batteries. J. Hazard. Mater. 150, 843-849 (2008).
- Gao, W. et al. Lithium carbonate recovery from cathode scrap of spent lithium-ion battery: a closed-loop process, Environ, Sci. Technol, 51, 1662-1669 (2017).
- Yang, Y. et al. A closed-loop process for selective metal recovery from spent lithium iron phosphate batteries through mechanochemical activation. ACS Sustain. Chem. Eng. 5, 9972-9980 (2017).
- Wang, M.-M., Zhang, C.-C. & Zhang, F.-S. An environmental benign process for cobalt and lithium recovery from spent lithium-ion batteries by mechanochemical approach. Waste Manag. 51, 239-244 (2016).
- 83. Wang, M.-M., Zhang, C.-C. & Zhang, F.-S. Recycling of spent lithium-ion battery with polyvinyl chloride by mechanochemical process. Waste Manag. 67, 232-239 (2017).
- Natarajan, S., Anantharaj, S., Tayade, R. J., Bajaj, H. C. & Kundu, S. Recovered spinel  $\mathsf{MnCo}_2\mathsf{O}_4$  from spent lithium-ion batteries for enhanced electrocatalytic oxygen evolution in alkaline medium. Dalton Trans. 46, 14382-14392 (2017).
- Xi, G., Zhao, T., Wang, L., Dun, C. & Zhang, Y. Effect of doping rare earths on magnetostriction characteristics of CoFe<sub>2</sub>O<sub>4</sub> prepared from spent Li-ion batteries. Physica B 534, 76-82 (2018).
- Moura, M. N. et al. Synthesis, characterization and photocatalytic properties of nanostructured CoFe<sub>2</sub>O<sub>4</sub> recycled from spent Li-ion batteries. Chemosphere 182.
- Li, J., Zhao, R., He, X. & Liu, H. Preparation of  $LiCoO_2$  cathode materials from spent lithium-ion batteries. Ionics 15, 111-113 (2009).
- Zou, H., Gratz, E., Apelian, D. & Wang, Y. A novel method to recycle mixed cathode materials for lithium ion batteries. Green Chem. 15, 1183-1191 (2013).
  - The process is elegantly designed to remove impurities and easily tunable to synthesize the current generation of cathode materials

# Review

- Sa, Q. et al. Synthesis of diverse LiNi<sub>x</sub>Mn<sub>v</sub>Co<sub>z</sub>O<sub>2</sub> cathode materials from lithium ion battery recovery stream. J. Sustain. Metall. 2, 248-256 (2016).
- Yang, Y., Xu, S. & He, Y. Lithium recycling and cathode material regeneration from acid leach liquor of spent lithium-ion battery via facile co-extraction and co-precipitation processes. Waste Manag. 64, 219-227 (2017).
- Li, L. et al. Sustainable recovery of cathode materials from spent lithium-ion batteries using lactic acid leaching system. ACS Sustain. Chem. Eng. 5, 5224-5233 (2017).
- 92. Liu, Y. & Liu, M. Reproduction of Li battery LiNi<sub>x</sub>Mn<sub>y</sub>Co<sub>1-x-y</sub>O<sub>2</sub> positive electrode material from the recycling of waste battery. Int. J. Hydrogen Energy 42, 18189-18195 (2017).
- Nithya, C., Thirunakaran, R., Sivashanmugam, A. & Gopukumar, S. High-performing LiMg<sub>x</sub>Cu<sub>y</sub>Co<sub>1-x-y</sub>O<sub>2</sub> cathode material for lithium rechargeable batteries. ACS Appl. Mater. Interfaces 4, 4040-4046 (2012).
- Shi, Y., Chen, G., Liu, F., Yue, X. & Chen, Z. Resolving the compositional and structural defects of degraded LiNi..Co..Mn.-O. particles to directly regenerate high-performance lithium-ion battery cathodes, ACS Energy Lett. 3, 1683-1692 (2018).
  - This paper highlights the importance of direct recycling to gain economic value from the resource.
- Dunn, J. B., Gaines, L., Sullivan, J. & Wang, M. Q. Impact of recycling on cradle-to-gate energy consumption and greenhouse gas emissions of automotive lithium-ion batteries. Environ, Sci. Technol. 46, 12704-12710 (2012).
  - This paper was one of the first to report the environmental burdens of material production, assembly and recycling of automotive LIBs in hybrid electric, plug-in hybrid electric, and battery electric vehicles.
- Sabisch, J. E. C., Anapolsky, A., Liu, G. & Minor, A. M. Evaluation of using pre-lithiated graphite from recycled Li-ion batteries for new LiB anodes. Resour. Conserv. Recycling 129, 129-134 (2018).
  - Whereas most papers focus on the recycling of valuable cathode materials, this examines the direct recycling of anode material
- Editorial. Recycle spent batteries. Nat. Energy 4, 253 (2019).
- Clemens, O. & Slater, P. R. Topochemical modifications of mixed metal oxide compounds by low-temperature fluorination routes. Rev. Inorg. Chem. 33, https://doi.org/10.1515/ revic-2013-0002 (2013).
- 99. Bolli, C., Guéguen, A., Mendez, M. A. & Berg, E. J. Operando monitoring of F formation in lithium ion batteries. Chem. Mater. 31, 1258-1267 (2019).
  - This paper suggests that the binder (PVDF) may also contribute to cell degradation and must be taken into account when developing future recycling methodologies
- 100. Karimi, G. R., Rowson, N. A. & Hewitt, C. J. Bioleaching of copper via iron oxidation from chalcopyrite at elevated temperatures, Food Bioprod, Process, 88, 21-25 (2010).
- 101. Smith, S. L., Grail, B. M. & Johnson, D. B. Reductive bioprocessing of cobalt-bearing limonitic laterites, Miner, Eng. 106, 86-90 (2017).
- 102. Horeh, N. B., Mousavi, S. M. & Shojaosadati, S. A. Bioleaching of valuable metals from spent lithium-ion mobile phone batteries using Aspergillus niger. J. Power Sources 320, 257-266 (2016).
- 103. Xin, Y. et al. Bioleaching of valuable metals Li, Co, Ni and Mn from spent electric vehicle Li-ion batteries for the purpose of recovery. J. Clean. Prod. 116, 249-258 (2016).

- 104. Mishra, D., Kim, D.-J., Ralph, D. E., Ahn, J.-G. & Rhee, Y.-H. Bioleaching of metals from spent lithium ion secondary batteries using Acidithiobacillus ferrooxidans. Waste Manag. 28, 333-338 (2008)
- 105. Pollmann, K., Raff, J., Merroun, M., Fahmy, K. & Selenska-Pobell, S. Metal binding by bacteria from uranium mining waste piles and its technological applications. Biotechnol. Adv. 24, 58-68 (2006).
- 106. Macaskie, L. E. et al. Today's wastes, tomorrow's materials for environmental protection. Hydrometallurgy 104, 483-487 (2010).
- 107. Ciez, R. E. & Whitacre, J. F. Examining different recycling processes for lithium-ion batteries, Nat. Sustain, 2, 148-156 (2019).

Acknowledgements Many of the ideas suggested for recovery of high-value materials will be trialled by the Faraday Institution's ReLiB fast-start project funded by the Faraday Institution (grant numbers FIRG005 and FIRG006) and by the ReCell Center, at Argonne National Laboratory, funded by the US Department of Energy. We acknowledge the contribution to the creation of the ReLiB project of N. Rowson (Birmingham Centre for Strategic Elements and Critical Materials). We also thank Q. Dai at Argonne National Laboratories for providing additional data for Fig. 6.

Author contributions G.H. and P.A. produced the original concept of the Review, and wrote the article, integrating contributions from the team and editing and shaping the review. G.H. produced the 'Social and environmental impacts of LIBs' section. R. Somerville and E.K. collaborated on the 'Physical materials separation' and 'Stabilization and passivation of end-oflife batteries' sections; E.K. produced the 'Biological recovery' section. L.D. and P.S. produced the 'Direct recycling' section and part of the 'Hydrometallurgical metals reclamation' section. R. Stolkin and A.W. collaboratively produced the 'Automating battery assembly' section. P.C. provided contributions on safety, and safe discharging of batteries, O.H. contributed to the supply and value chain, environmental impact and economic assessments and S.L. provided information on battery re-use. A.A. and K.R. produced most of the 'Hydrometallurgical metals reclamation' section. L.G. critically revised the article. Figures 1 and 2 were created by G.H. (with help from R. Somerville and E.K.) and Fig. 4 was created by R. Somerville. Figure 3 was created by L.D., P.A. and G.H. and Fig. 6 was created by G.H. and L.G.

Competing interests The authors declare no competing interests.

#### Additional information

Correspondence and requests for materials should be addressed to G.H. or P.A. Peer review information Nature thanks Anand Bhatt and Matthew Lacey and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.  $\textbf{Reprints and permissions information} \ is \ available \ at \ http://www.nature.com/reprints.$ Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© Springer Nature Limited 2019

# The technological and economic prospects for CO<sub>2</sub> utilization and removal

https://doi.org/10.1038/s41586-019-1681-6

Received: 24 September 2018

Accepted: 13 September 2019

Published online: 6 November 2019

Cameron Hepburn<sup>1,2</sup>, Ella Adlen<sup>1\*</sup>, John Beddington<sup>1</sup>, Emily A. Carter<sup>3,4,5</sup>, Sabine Fuss<sup>6,7</sup>, Niall Mac Dowell<sup>8</sup>, Jan C. Minx<sup>6,9</sup>, Pete Smith<sup>10</sup> & Charlotte K. Williams<sup>11</sup>

The capture and use of carbon dioxide to create valuable products might lower the net costs of reducing emissions or removing carbon dioxide from the atmosphere. Here we review ten pathways for the utilization of carbon dioxide. Pathways that involve chemicals, fuels and microalgae might reduce emissions of carbon dioxide but have limited potential for its removal, whereas pathways that involve construction materials can both utilize and remove carbon dioxide. Land-based pathways can increase agricultural output and remove carbon dioxide. Our assessment suggests that each pathway could scale to over 0.5 gigatonnes of carbon dioxide utilization annually. However, barriers to implementation remain substantial and resource constraints prevent the simultaneous deployment of all pathways.



CO<sub>2</sub> utilization is receiving increasing interest from the scientific community<sup>1</sup>. This is partly due to climate change considerations and partly because using CO<sub>2</sub> as a feedstock can result in a cheaper or cleaner production process compared with using conventional hydrocarbons<sup>2</sup>. CO<sub>2</sub> utilization is often promoted as a way to reduce the net costs—or increase the profits—of reducing emissions or removing carbon dioxide from the atmosphere, and therefore as a way to aid the scaling of mitigation or removal efforts<sup>3</sup>. CO<sub>2</sub> utilization is also seen variously as a stepping stone towards<sup>4</sup> or a distraction away from<sup>5</sup> the successful implementation of carbon capture and storage (CCS) at scale.

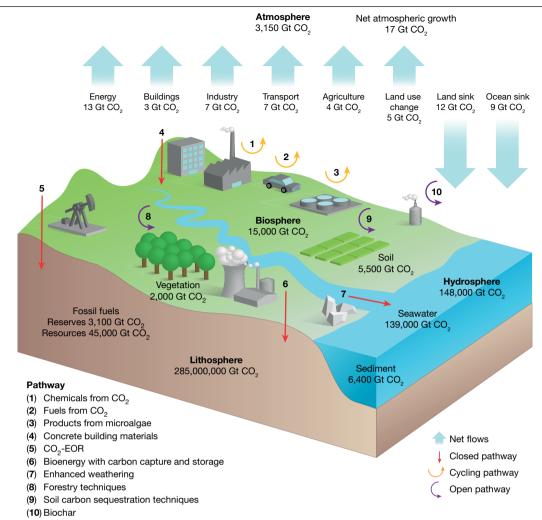
In most of the literature—including the IPCC 2005 Special Report on Carbon Dioxide Capture and Storage<sup>6</sup>—the term 'CO<sub>2</sub> utilization' refers to the use of CO<sub>2</sub>, at concentrations above atmospheric levels, directly or as a feedstock in industrial or chemical processes, to produce valuable carbon-containing products<sup>6-11</sup>. Included in this conventional definition is the industrial production of fuels using, for example, amines to capture and concentrate the CO<sub>2</sub> from air, potentially with solar energy. However, the definition excludes cases in which an identical fuel is produced from the same essential inputs, but the CO<sub>2</sub> utilized is captured by plant-based photosynthetic processes.

Here, we consider CO<sub>2</sub> utilization to be a process in which one or more economically valuable products are produced using CO<sub>2</sub>, whether the  $CO_2$  is supplied from fossil-derived waste gases, captured from the atmosphere by an industrial process, or-in a departure from most (but not all<sup>12,13</sup>) of the literature—captured biologically by land-based processes. Biological or land-based forms of CO<sub>2</sub> utilization can generate economic value in the form of, for example, wood products for buildings, increased plant yields from enhanced soil carbon uptake, and even the production of biofuel and bio-derived chemicals. We use this broader definition deliberately; by thinking functionally, rather than narrowly about specific processes, we hope to promote dialogue across scientific fields, compare costs and benefits across pathways, and consider common techno-economic characteristics across pathways that could potentially assist in the identification of routes towards the mitigation of climate change.

In this Perspective, we consider a non-exhaustive selection of ten CO<sub>2</sub> utilization pathways and provide a transparent assessment of the potential scale and cost for each one. The ten pathways are as follows: (1) CO<sub>2</sub>-based chemical products, including polymers: (2) CO<sub>2</sub>-based fuels; (3) microalgae fuels and other microalgae products; (4) concrete building materials; (5) CO<sub>2</sub> enhanced oil recovery (CO<sub>2</sub>-EOR); (6) bioenergy with carbon capture and storage (BECCS); (7) enhanced weathering; (8) forestry techniques, including afforestation/reforestation, forest management and wood products; (9) land management via soil carbon sequestration techniques; and (10) biochar.

These ten CO<sub>2</sub> utilization pathways can also be characterized as 'cycling', 'closed' and 'open' utilization pathways (Fig. 1, Table 1, Supplementary Materials). For instance, many (but not all) conventional industrial utilization pathways-such as CO<sub>2</sub>-based fuels and chemicals—tend to be 'cycling': they move carbon through industrial systems over timescales of days, weeks or months. Such pathways do not provide net CO<sub>2</sub> removal from the atmosphere, but they can reduce emissions via industrial CO<sub>2</sub> capture that displaces fossil fuel use. By contrast, 'closed' pathways involve utilization and nearpermanent  $CO_2$  storage, such as in the lithosphere (via  $CO_2$ -EOR or BECCS), in the deep ocean (via terrestrial enhanced weathering) or in mineralized carbon in the built and natural environments. Finally, 'open' pathways tend to be based in biological systems,

10xford Martin School, University of Oxford, Oxford, UK. 2Smith School of Enterprise and the Environment, University of Oxford, Oxford, UK. 3School of Engineering and Applied Science, Princeton University, Princeton, NJ, USA. <sup>4</sup>Department of Chemical and Biomolecular Engineering, University of California, Los Angeles, CA, USA. <sup>5</sup>Office of the Chancellor, University of California, Los Angeles, CA, USA. Mercator Research Institute on Global Commons and Climate Change, Berlin, Germany. Department of Geography, Humboldt University of Berlin, Berlin, Germany. <sup>8</sup>Centre for Environmental Policy, Imperial College London, London, UK. <sup>9</sup>School of Earth and Environment, University of Leeds, Leeds, UK. <sup>10</sup>Institute of Biological and Environmental Sciences, University of Aberdeen, Aberdeen, UK, 11 Department of Chemistry, University of Oxford, Oxford, UK, \*e-mail: ella.adlen@oxfordmartin.ox,ac.uk



 $\label{eq:converted} \textbf{Fig. 1}| \textbf{Stocks and net flows of CO}_2 \textbf{ including potential utilization and removal pathways.} \text{ } \textbf{O} \textbf{range}, \textbf{red and purple arrows (numbered 1-10, as described in Table 1) represent cycling, closed and open pathways for CO}_2 \textbf{ utilization and removal. Teal block arrows represent annual flows to and from the atmosphere, with estimates averaged over the 2008-2017 period^{15,91}. \\ \textbf{Estimates of stocks in the Earth's spheres (lithosphere, biosphere, hydrosphere and atmosphere, labelled in bold) and selected stock subcategories are given. \\ \textbf{All estimates are based on IPCC estimates}^{16} \textbf{ except where noted, and are converted from C to CO}_2. Carbon stocks in the hydrosphere comprise seawater,} \\ \end{aligned}$ 

sediment, and dissolved organic carbon (not shown, around 2,600 Gt CO<sub>2</sub>). The vast majority of carbon stocks in the lithosphere are locked in the Earth's crust  $^{92}$ , with estimated accessible fossil fuel reserves and resources of more than  $45,000~\rm Gt~\rm CO_2^{25}$ . Atmospheric stocks are converted from the 2017 estimates of atmospheric CO $_2$  of  $405~\rm ppm^{93}$  using a conversion factor of 2.12. Carbon stocks in the biosphere include those stored in permafrost and wetlands (not shown, around 7,500 Gt CO $_2$ ), vegetation, and soils. Soil stocks to 1-m depth have been recently estimated at 5,500 Gt CO $_2^{62}$ .

and are characterized by large removal potentials and storage in 'leaky' natural systems—such as biomass and soil—with the risk of large-scale flux back to the atmosphere.

Of the pathways we discuss, some are novel or emerging—such as  $\mathrm{CO}_2$ -fuels, for which current flows are near-zero—whereas others are well established, such as  $\mathrm{CO}_2$ -EOR and afforestation/reforestation. Pathways were selected on the basis of discussions at a joint meeting of the US National Academy of Sciences and the UK Royal Society<sup>1</sup>; each pathway is relatively well studied to date and has an acknowledged potential to scale. There are many other pathways that meet our definition but are not reviewed here (Supplementary Materials).

This Perspective is structured as follows: first, the ten utilization pathways are presented in the context of the scale of  $\mathrm{CO}_2$  stocks and flows on Earth. Second, the potential scale and economics of each pathway are assessed. Third, a selection of key barriers to scaling is identified. Fourth, we assess the outlook for  $\mathrm{CO}_2$  utilization, and conclude with priorities for future research and policy.

# CO<sub>2</sub> utilization and the carbon cycle

The amount of carbon dioxide that is utilized by a pathway is not necessarily the same as the amount of carbon dioxide removed or carbon dioxide stored.  $CO_2$  utilization does not necessarily reduce emissions and does not necessarily deliver a net climate benefit, once indirect and other effects have been accounted for. The various concepts overlap and relate to each other, but are distinct (Supplementary Fig. 1, Supplementary Materials). Some carbon capture and utilization (CCU) processes achieve carbon dioxide removal (CDR) from the atmosphere, and some involve CCS. CCS itself can contribute either to the mitigation of  $CO_2$  (for example, by reducing net emissions from a gas-fired power plant) or to atmospheric removals (for example, by direct air carbon capture and storage, or DACCS); CCS does not necessarily imply CDR. Furthermore, CCS and CDR can fail to deliver a climate benefit. For instance, perverse indirect effects—such as land-use change resulting from BECCS<sup>14</sup>—could increase net atmospheric  $CO_2$  concentrations.

Table 1 | Ten CO<sub>2</sub> utilization and removal pathways

Pathway <sup>a</sup>	Removal and/or capture <sup>b</sup>	Utilization product	Storage <sup>c,d</sup> and likelihood of release (high/low)	Emission on use <sup>f</sup> or release during storage <sup>g</sup>	Example cycles <sup>h</sup>
(1) Chemicals from CO <sub>2</sub>	Catalytic chemical conversion of $\mathrm{CO}_2$ from flue gas or other sources into chemical products	CO <sub>2</sub> -derived platform chemicals such as methanol, urea and plastics	Various chemicals (days/ decades) – high	Hydrolysis or decomposition	KCLG; KCLF; ALFJ; ALG
(2) Fuels from CO <sub>2</sub>	Catalytic hydrogenation processes to convert CO <sub>2</sub> from flue gas or other sources into fuels	CO <sub>2</sub> -derived fuels such as methanol, methane and Fischer–Tropschderived fuels  Various fuels (weeks/ Combustion months) – high		Combustion	KCLG; ALG
(3) Products from microalgae	Uptake of CO <sub>2</sub> from the atmosphere or other sources by microalgae biomass	mosphere or other sources by bioproducts such as months) - high or consu		Combustion (fuel) or consumption (bioproduct)	KCLG; BG
(4) Concrete building materials	Mineralization of CO <sub>2</sub> from flue gas or other sources into industrial waste materials, and CO <sub>2</sub> curing of concrete	Carbonated aggregates Carbonates (centuries) Extreme acid conditions		KCLF; ALF	
(5) CO <sub>2</sub> -EOR	Injection of CO <sub>2</sub> from flue gas or other sources into oil reservoirs	Oil	Geological sequestration (millennia) – low <sup>e</sup>	n.a.	KCD
(6) Bioenergy with carbon capture and storage (BECCS)	Growth of plant biomass	Bioenergy crop biomass	Geological sequestration (millennia) – low <sup>e</sup>	n.a.	BCD
(7) Enhanced weathering	Mineralization of atmospheric CO <sub>2</sub> via the application of pulverized silicate rock to cropland, grassland and forests	Agricultural crop biomass	Aqueous carbonate (centuries) – low	Extreme acidic conditions	BE
(8) Forestry techniques	Growth of woody biomass via afforestation, reforestation or sustainable forest management	Standing biomass, wood products	Standing forests and long-lived wood products (decades to centuries) – high	Disturbance, combustion or decomposition	BFJ
(9) Soil carbon sequestration techniques	Increase in soil organic carbon content via various land management practices	Agricultural crop biomass	Soil organic carbon (years to decades) – high	Disturbance or decomposition	BFJ
(10) Biochar	Growth of plant biomass for pyrolysis and application of char to soils	Agricultural or bioenergy crop biomass	Black carbon (years to decades) – high	Decomposition	BFJ
n a not applicable					

CO<sub>2</sub> utilization does not necessarily contribute to addressing climate change, and careful analysis is essential to determine its overall impact. Identifying the counterfactual—what would have happened without CO<sub>2</sub> utilization—is important but is often particularly challenging, and the impact of a given CO<sub>2</sub> utilization pathway on the mitigation of climate change varies as a function of space and time (Box 1).

 $For \, CO_2 utilization \, to \, contribute \, usefully \, to \, the \, reduction \, of \, atmos$ pheric CO<sub>2</sub> concentrations, the scale of the pathways must be meaningful in comparison with the net flows of CO<sub>2</sub> shown in Fig. 1. The flux of carbon from fossil fuels and industry to the atmosphere (34 Gt CO<sub>2</sub> yr<sup>-1</sup>)<sup>15</sup> is dwarfed by the gross flux to land via photosynthesis in plants  $(440 \,\mathrm{Gt}\,\mathrm{CO}_2\,\mathrm{yr}^{-1})^{16}$ . However, only 2%-3% of this photosynthetic carbon remains on land (12 Gt CO<sub>2</sub> yr<sup>-1</sup>), and only for decades; the remainder is re-emitted by plant and soil respiration. If soil carbon uptake could be increased by 0.4% per year, this would contribute to achieving net zero emissions—as per the '4 per mille' initiative<sup>17</sup>—but this is challenging<sup>18</sup>. Of the ten pathways we discuss, five leverage our ability to perturb these land-based fluxes.

The other five conventional industrial CO<sub>2</sub> utilization pathways could also perturb the net flows of CO<sub>2</sub>. The production of plastics and other products creates a demand for so-called 'socioeconomic carbon' 19 (around 2.4 Gt CO<sub>2</sub> yr<sup>-1</sup>, of which around two-thirds is wood products) that could be met in part through CO<sub>2</sub> utilization. The total stock of carbon accumulated in products (such as wood products, bitumen, plastic and cereals) has been estimated at 42 Gt CO<sub>2</sub> in 2008, of which 25 Gt CO<sub>2</sub> is in wood products<sup>19</sup>. Up to 16 Gt CO<sub>2</sub> was sequestered in human infrastructure as mineralized carbonates in cement between  $1930\,and\,2013, with\,current\,rates^{20,21}\,estimated\,to\,be\,around\,1\,Gt\,CO_2\,yr^{-1}.$ 

The flow of CO<sub>2</sub> through the different utilization pathways can be represented by a combination of different steps (labels A to L; Fig. 2, Table 1). Utilization pathways often (but not always) involve removal (A or B) and storage (D, E or F); however, the permanence of CO<sub>2</sub> storage varies greatly from one utilization pathway to another, with storage timeframes ranging from days to millennia. In part, permanence depends upon where the carbon ends up (Fig. 1): the lithosphere, by geological sequestration into reservoirs such as saline aquifers or

The ten pathways are depicted in Fig. 1 and are represented as a combination of steps in Fig. 2.

<sup>&</sup>lt;sup>b</sup>Removal and/or capture corresponds to steps A, B and/or C in Fig. 2.

<sup>°</sup>Storage corresponds to steps D, E or F in Fig. 2.

dStorage durations represent best-case scenarios. For instance, in CO.-EOR, if the well is operated with complete recycle, the CO. is trapped and can be stored on a timescale of centuries or more<sup>22</sup>. This is also relevant only for conventional operations

eRelease during geological storage is usually a consequence of engineering implementation error.

<sup>&</sup>lt;sup>f</sup>Emission on use corresponds to step G in Fig. 2.

<sup>&</sup>lt;sup>9</sup>Release during storage corresponds to steps H, I or J in Fig. 2.

<sup>&</sup>lt;sup>h</sup>The letters stated are the steps from Fig. 2 that comprise the example cycle.

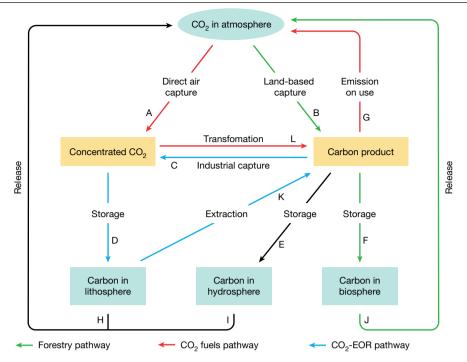


Fig. 2 | Carbon dioxide utilization and removal cycle. Utilization pathways are represented as a combination of steps, A-L. Green arrows trace an example open pathway, forestry (BFJ). Red arrows trace an example cycling pathway, CO<sub>2</sub> fuels with direct air capture (ALG). Blue arrows trace an example closed pathway,

CO<sub>2</sub>-EOR (KCD). Cycling pathways (with the exception of polymers) end with step G; closed pathways end with steps D, E or F; and open pathways end with step J. See Table 1 for further description. All flows are net of process emissions.

depleted oil and gas reservoirs, or by mineralization into rocks; the biosphere, in trees, soils and the human-built environment; or the hydrosphere, with storage in the deep oceans. Geological storage, when executed correctly, is considered to be more permanent<sup>22</sup> than storage in the biosphere, which is shorter and subject to human and natural disturbances<sup>23</sup> such as wildfires and pests, as well as changes in climate<sup>24</sup>. However, even 'closed' pathways do not offer completely permanent storage over geological timescales (more than 100,000 years<sup>25</sup>), which gives rise to intergenerational ethical questions<sup>26</sup>.

In the short term, the creation of products from concentrated CO<sub>2</sub>, as in step L (albeit, CO<sub>2</sub> conversion is not a necessary requirement for utilization), could leverage the industrial capture of flue gases following the extraction and combustion of fossil fuels (KC)<sup>27</sup>. In the longer term, the CO<sub>2</sub> loop will need to be closed in order to achieve net zero emissions. implying that CO<sub>2</sub> will need to be sourced from the atmosphere, potentially via direct air capture (A) or through land-based uptake by photosynthesis or mineralization (B). For instance, net zero CO<sub>2</sub>-based fuels must shift the current flows of carbon, from a lithosphere-to-atmosphere (KCLG) to an atmosphere-to-atmosphere cycle (ALG) (Fig. 2).

# Scale and economics of CO<sub>2</sub> utilization

We assess the peer-reviewed literature on the ten pathways, which comprises over 11,000 papers. For the conventional pathways, our scoping review covered over 5,000 papers, a minority (186) of which provide cost estimates. Estimates of potential scale were informed by a structured estimation process and an expert opinion survey. For the non-conventional utilization pathways, we build upon existing CO<sub>2</sub> removal estimates (also derived from a scoping review<sup>28</sup> of over 6,000 papers—of which 927 provide usable estimates—and an expert judgement process) and identify preliminary published research on the relationship between CO<sub>2</sub> removal and CO<sub>2</sub> utilization to offer estimates of the scale and cost of CO<sub>2</sub> utilization.

Where possible, we calculate breakeven costs in 2015 US dollars per tonne of CO<sub>2</sub> for each pathway (hereafter, all costs stated are in US dollars). The breakeven CO<sub>2</sub> cost represents the incentive per tonne of CO<sub>2</sub> utilized that would be necessary to make the pathway economic (see Supplementary Materials, S1.2). This can be thought of as the breakeven (theoretical) subsidy per tonne of CO<sub>2</sub> utilization, although we are not recommending such a subsidy.

#### **Conventional utilization pathways**

Dependent on a multitude of technological, policy and economic factors that remain unresolved, each of the conventional pathways-chemicals, fuels, microalgae, building materials and CO<sub>2</sub>-EOR-might utilize around 0.5 Gt CO<sub>2</sub> vr<sup>-1</sup> or more in 2050. We also estimate that between 0.2 and  $3.2\,Gt\,CO_2\,yr^{-1}$  could be removed and stored in the lithosphere or in the biosphere for centuries or more.

Chemicals. CO<sub>2</sub> can be transformed efficiently into a range of chemicals, but only a few of the technologies are economically viable and scalable. Some are commercialized<sup>29</sup>, such as the production of urea<sup>30</sup> and polycarbonate polyols<sup>31</sup>. Some are technically possible but are not widely adopted, such as the production of CO<sub>2</sub>-derived methanol in the absence of carbon monoxide<sup>32</sup> (methanol is a platform chemical for a multitude of other reaction pathways, including to fuels, and is mainly manufactured via the hydrogenation of a mixture of CO and 1%–2% CO<sub>2</sub>). Breakeven costs per tonne of CO<sub>2</sub>, calculated from the scoping review, for urea (around -\$100) and for polyols (around -\$2,600) reflect that these markets are currently profitable. The estimated utilization potential for CO<sub>2</sub> in chemicals is around 0.3 to 0.6 Gt CO<sub>2</sub> yr<sup>-1</sup> in 2050, and the interquartile range of breakeven costs obtained from the scoping review is -\$80 to \$320 per tonne of CO<sub>2</sub>.

Currently, the largest-scale chemical utilization pathway is that of urea production. 140 Mt CO<sub>2</sub> yr<sup>-1</sup> is utilized to produce 200 Mt yr<sup>-1</sup> of urea 33. Urea is produced from ammonia (which is generated by the energy-intensive Haber-Bosch process; 3H<sub>2</sub> + N<sub>2</sub> → 2NH<sub>3</sub>) and CO<sub>2</sub> according to  $2NH_3 + CO_2 \neq CO(NH_2)_2 + H_2O$ ; coal or natural gas typically provides the necessary energy. Within days of being applied as fertilizer, the carbon in urea is released to the atmosphere. For urea to

# Box 1

# CO<sub>2</sub> utilization, removal, storage, reduced emissions and net climate benefit

Does CO<sub>2</sub> utilization (CO<sub>2</sub>u) lead to a climate benefit? It might reduce emissions ( $CO_2\rho$ ), or remove  $CO_2$  ( $CO_2r$ ) from the atmosphere, and/or store it (CO2s). But various direct and indirect effects over the relevant life cycle must be considered and compared to a plausible baseline or 'counterfactual'—what would have happened without CO<sub>2</sub> utilization<sup>83</sup>. Assiduously calculating direct impacts in one place, and at one time, is of little use if there is a 'waterbed effect' (also referred to as a 'rebound' or 'leakage') in which emissions occur somewhere else, or later.

For instance, obtaining a barrel of oil via CO<sub>2</sub>-EOR utilizes CO<sub>2</sub>, which can remain in the oil formation rather than being re-emitted into the atmosphere. Assuming that the CO2 does not return to atmosphere, the CO2 utilized is equal to the CO2 emissions stored, that is,  $CO_2u = CO_2s$ , but whether  $CO_2r \ge 0$  depends upon the source of the CO2; if it is from a fossil power station, there is no net removal of CO<sub>2</sub> from the atmosphere. Emissions have been reduced, and  $CO_2\rho = CO_2u = CO_2s > 0$ , even though  $CO_2r = 0$ .

To visualize this, consider a 'reference' scenario in which 1 t CO<sub>2</sub> is emitted from a fossil power plant, and 1.5 t CO<sub>2</sub> are emitted from oil use, such that total emissions are 2.5 t CO<sub>2</sub>. Compare this to a 'utilization' scenario, in which the CO<sub>2</sub> from the power plant is used for CO<sub>2</sub>-EOR instead—that is, CO<sub>2</sub>u = 1 t CO<sub>2</sub>. Total emissions in this 'utilization' scenario comprise the 1.5 t CO<sub>2</sub> from the consumption of the CO<sub>2</sub>-EOR oil. Emissions reduced is equal to  $2.5 - 1.5 = 1.0 \text{ t CO}_2\rho$ , which is identical to the CO<sub>2</sub>u, but net  $CO_2$ r = 0 because the  $CO_2$  came from a fossil power plant, rather than from the atmosphere.

In reality, the emissions from the baseline barrel of oil that was displaced by the CO<sub>2</sub>-EOR oil might be higher or lower, depending on its origin and its production process. If the CO<sub>2</sub>-EOR oil displaces the use of renewable electricity in an electric vehicle, CO<sub>2</sub>-EOR generates a net increase in emissions. If CO<sub>2</sub>-EOR is to offer net removals, the CO<sub>2</sub> must be captured from the atmosphere, and more carbon must be injected into the well than is extracted.

Life-cycle analyses on some industrial CO<sub>2</sub> utilization pathways suggest that the potential for net emission reductions is much larger than for net removals, which appears very modest<sup>94</sup>. Up to 3 tonnes of CO<sub>2</sub> emissions may be avoided for every 1 tonne of CO<sub>2</sub> used in polycarbonate polyols<sup>2</sup>, even though no CO<sub>2</sub> is removed from atmosphere. Nearly 4 tonnes of CO<sub>2</sub> emissions may be avoided for each tonne of dry wood used that displaces concrete-based materials95.

Other life-cycle analyses have found neutral or negative impacts of CO<sub>2</sub> utilization on reducing emissions<sup>74,96-98</sup>. For instance, CO<sub>2</sub> utilization pathways that involve the input of energy from nondecarbonized sources may result in net life-cycle increases in CO<sub>2</sub>96-99.

be net zero carbon, it would require its carbon to be sourced from the atmosphere-for example, using direct air capture-and the energy source would need to be renewable. All nitrogen-based fertilizers produce N<sub>2</sub>O, a greenhouse gas that is around 300 times more potent than CO<sub>2</sub> over a 100-year time horizon<sup>34</sup>. Increasing urea production may therefore have a negative impact on climate<sup>35</sup>.

For the production of polymers, the utilization potential of CO<sub>2</sub> is estimated to be 10 to 50 Mt vr<sup>-1</sup> in 2050. In the current market structure. around 60% of plastics have applications in sectors other than packaging-including as durable materials for construction, household goods, electronics, and in vehicles. Such products have lifespans of decades or even centuries36.

Fuels and microalgae. Fuels derived from CO<sub>2</sub> are argued to be an attractive option in the decarbonization process<sup>37,38</sup> because they can be deployed within existing transport infrastructure. Such fuels could also find a role in sectors that are harder to decarbonize, such as aviation<sup>39</sup>, since hydrocarbons have energy densities that are orders of magnitude above those of present-day batteries<sup>32</sup>. The long-term use of carbonbased energy carriers in a net zero emissions economy relies upon their production with renewable energy, and upon low-cost, scalable, clean hydrogen production-for example via the electrolysis of water or by novel alternative methods.

Here we consider products such as methanol, methane, dimethyl ether, and Fischer-Tropsch fuels as potential CO2 energy carriers for transportation. The estimated potential for the scale of  $CO_2$  utilization in fuels varies widely, from 1 to 4.2 Gt CO<sub>2</sub> yr<sup>-1</sup>, reflecting uncertainties in potential market penetration. The high end represents a future in which synfuels have sizeable market shares, due to cost reductions and policy drivers. The low end-which is itself considerable-represents very modest penetration into the methane and fuels markets, but it could also be an overestimate if CO<sub>2</sub>-derived products do not become costcompetitive with alternative clean energy vectors such as hydrogen or ammonia, or with direct sequestration.

ACO<sub>2</sub>-to-methanol plant operates in Iceland, and various power-togas plants operate worldwide. However, these plants represent special cases that may be difficult to replicate because they are exploiting geographic advantages, such as the availability of cheap geothermal energy.  $Although the production of more complex \ hydrocarbons \ is energetically$ and therefore economically expensive11, rapid cost-reductions could potentially occur if renewable energy—which represents a large proportion of total cost-continues to become cheaper, and if policy stimulates other cost reductions. The US Department of Energy's target for the cost of hydrogen production—\$2 per kg of H<sub>2</sub>—is roughly equivalent to \$2 per gasoline-gallon equivalent, and would require carbon-free electricity to cost less than \$0.03 kWh<sup>-1</sup> (accounting for kinetics and other losses to the enthalpy of electrolysis-based hydrogen production, around 40 kWh per kg H<sub>2</sub>)<sup>40</sup>. In recent years, several wind and solar power auctions around the world have been won with prices below 41 \$0.03 kWh<sup>-1</sup>.

The interquartile range for breakeven costs for CO<sub>2</sub> fuels from our scoping review was \$0 to \$670 per tonne of CO<sub>2</sub>. Negative breakeven costs appear in studies that model particularly beneficial scenarios, such as low discount rates, free feedstocks, or free or low-cost renewable electricity.

For pathways that have high capital costs, the benefits of economies of scale and learning could be considerable 42. This is particularly relevant for the algal pathways that require photobioreactors<sup>43</sup> and for the fuel synthesis pathways that require electrolysers<sup>44</sup>. Microalgae are a subject of long-standing research interest because of their high CO<sub>2</sub>-fixation efficiencies (up to 10%, compared with 1%-4% for other biomass<sup>45</sup>), as well as their potential to produce a range of products such as biofuels, high-value carbohydrates and proteins, and plastics<sup>43</sup>. The microalgae pathway has complex production economics and the estimated CO<sub>2</sub> utilization potential for microalgae in 2050 ranges from 0.2 to 0.9 Gt CO<sub>2</sub> yr<sup>-1</sup>, with a breakeven cost interquartile range from the scoping review of \$230 to \$920 per tonne of CO<sub>2</sub>.

Concrete building materials. CO<sub>2</sub> utilization pathways in concrete building materials are estimated to remove, utilize and store between 0.1 and 1.4 Gt CO<sub>2</sub> yr<sup>-1</sup> over the long term—with the CO<sub>2</sub> sequestered well beyond the lifespan of the infrastructure itself-at interquartile

Table 2 | Range estimates of the potential for CO2 utilization and present-day breakeven cost

Pathway	Removal potential in 2050 (Mt $CO_2$ removed per year)	Utilization potential in 2050 (Mt $CO_2$ utilized per year)	Breakeven cost of $CO_2$ utilization (2015 US\$ per tonne $CO_2$ utilized)
Conventional utilization			
Chemicals	Around 10 to 30	300 to 600	-\$80 to \$320
Fuels	0	1,000 to 4,200	\$0 to \$670
Microalgae	0	200 to 900	\$230 to \$920
Concrete building materials	100 to 1,400	100 to 1,400	-\$30 to \$70
Enhanced oil recovery	100 to 1,800	100 to 1,800	-\$60 to -\$45
Non-conventional utilization			
BECCS	500 to 5,000	500 to 5,000	\$60 to \$160
Enhanced weathering	2,000 to 4,000	n.d.	Less than \$200*
Forestry techniques	500 to 3,600	70 to 1,100	-\$40 to \$10
Land management	2,300 to 5,300	900 to 1,900	-\$90 to -\$20
Biochar	300 to 2,000	170 to 1,000	-\$70 to -\$60

n.d., not determined.

The breakeven cost is the cost in 2015 US\$ per tonne of CO<sub>2</sub> adjusted for revenues, by-products, and any CO<sub>2</sub> credits or fees. A breakeven cost of zero represents the point at which the pathway is economically viable without governmental CO<sub>2</sub> pricing (for example, a subsidy for CO<sub>2</sub> utilization). Breakeven costs presented as a range represent either (for conventional pathways with the exception of EOR) 25th and 75th percentile estimates as calculated via the scoping review of the academic literature (in which the magnitude of the difference reflects the diversity of technological and economic assumptions available within and across each sub-pathway) or (for land-based pathways) top-down estimates of revenues that may accrue (when the uncertainty of the accuracy of the estimation is high). Breakeven costs presented with an asterisk are calculated unadjusted for revenues and by-product credits. To obtain the global gross utilization potential high and low values for conventional pathways, we averaged the interpolated expert opinions with an author group estimate. For non-conventional utilization pathways, estimated utilization potential ranges are based on estimates of additional realized yield of carbon in vegetation (for soil carbon sequestration and biochar, additional yield approximates to net primary productivity, and for afforestation/reforestation, it approximates to wood products). These are first rough estimates based on preliminary but sparse published research reporting relationships between carbon storage and additional carbon that can be utilized.

breakeven costs of -\$30 to \$70 per tonne of  $CO_2$ . The high end might reflect a scenario (amongst other possibilities) in which  $CO_2$  is used as a cement curing agent in the entirety of the precast concrete market and in 70% of the pourable cement markets. The estimate also includes aggregates that are produced from carbonated industrial wastes, such as cement and demolition waste, steel slag, cement kiln dust, and coal pulverized fuel ash.

Cement requires the use of lime (CaO), which is produced by the calcination of limestone in an emissions-intensive process. As such, unless calcination is paired with carbon capture and sequestration, it is difficult for building-related pathways to deliver reductions in  $CO_2$  emissions on a life-cycle basis. Several commercial initiatives aim to replace the lime-based ordinary Portland cement—which currently dominates the global market—with alternative binders such as steel-slag based systems <sup>46</sup> or geopolymers made from aluminosilicates <sup>47</sup>.

 ${\bf CO_2\text{-}EOR}$ . Enhanced oil recovery using  ${\rm CO_2}$  currently accounts for around 5% of the total US crude oil production<sup>48</sup>. Conventionally, operators aim to maximize both the amount of oil recovered and the amount of  ${\rm CO_2}$  recovered (rather than  ${\rm CO_2}$  stored) per tonne of  ${\rm CO_2}$  injected; between 1.1 and 3.3 barrels (bbl) of oil can be produced per tonne of  ${\rm CO_2}$  injected under conventional operation and within the constraints of natural reservoir heterogeneity<sup>49</sup>. However, in principle—and depending on operating conditions and project type— ${\rm CO_2\text{-}EOR}$  can be operated such that, on a life-cycle basis, more  ${\rm CO_2}$  is injected than is produced upon consumption of the final oil product<sup>50</sup>.

More than 90% of the world's oil reservoirs are potentially suitable for  $CO_2$ -EOR<sup>51</sup>, which implies that as much as 140 Gt  $CO_2$  could be used and stored in this way<sup>5</sup>. We estimate a 2050 utilization rate of around 0.1 to 1.8 Gt  $CO_2$  yr<sup>-1</sup>. If EOR was deployed to maximize  $CO_2$  storage—rather than oil output—then genuine  $CO_2$  emission reductions are possible, depending on the emissions intensity of the counterfactual and on the relevant inefficiencies (Box 1).

At oil prices of approximately  $100 \, \text{bbl}^{-1}$ , EOR is economically viable if CO<sub>2</sub> can be sourced for between  $45 \, \text{and} \, 60 \, \text{per tonne}$  of CO<sub>2</sub>  $^{49,51}$ ,

implying a breakeven cost of  $CO_2$  of -\$60 to -\$45 per tonne of  $CO_2$ . These cost estimates (realistically or unrealistically) assume \$100 bbl $^{-1}$  oil prices and are specific to the United States, where the business model is mature.

## Non-conventional utilization pathways

The five non-conventional utilization pathways that we review here are BECCS, enhanced weathering, forestry techniques, land management practices, and biochar. Previous reviews  $^{18,28,52-54}$  have shown that these pathways offer substantial CO $_2$  removal potential: a recent substantive scoping review  $^{28}$  gives values of 0.5 to 3.6 Gt CO $_2$  yr  $^{-1}$  for afforestation/reforestation, 2.3 to 5.3 Gt CO $_2$  yr  $^{-1}$  for land management, 0.3 to 2 Gt CO $_2$  yr  $^{-1}$  for biochar, and 0.5 to 5 Gt CO $_2$  yr  $^{-1}$  for BECCS. Enhanced weathering offers a removal potential of 2 to 4 Gt CO $_2$  yr  $^{-1}$  at costs  $^{28}$  of around \$200 per tonne of CO $_2$ . Not all of this potential involves utilization of carbon dioxide resulting in economic value, but the approximate scale of CO $_2$  utilized that is described below could be considerable. The breakeven costs per tonne of CO $_2$  utilized that we estimate here are low and are frequently negative.

**BECCS**. BECCS involves the biological capture of atmospheric carbon by photosynthetic processes, producing biomass used for the generation of electricity or fuel, before  $CO_2$  is captured and removed. Although there is substantial uncertainty regarding the total quantity of available biomass<sup>55</sup>—particularly in light of concerns over competition for land use with food crops—100 to 300 EJ yr $^{-1}$  of primary energy equivalent of biomass could be deployed by 2050.

BECCS provides two distinct services: bioenergy, and atmospheric  $\mathrm{CO}_2$  removal. Although several cost estimates exist in the literature—for example, around \$200 per tonne of  $\mathrm{CO}_2^{28}$ —these typically assign all costs to the  $\mathrm{CO}_2$  removal service, and thus implicitly assume that no revenue is received for the bioenergy services that are generated. By approximating those revenues using a basket of wholesale electricity prices across countries that are suited to host BECCS systems <sup>56</sup>, we estimate breakeven costs of between \$60 and \$160 per tonne of  $\mathrm{CO}_2$  utilized.

Table 3 | Costs of utilization compared with product costs, scoping review

Pathway	Cost of product made with CO <sub>2</sub> utilization (US\$ per tonne of product) Median, scoping review	Selling price of product (US\$ per tonne of product) Present day	Difference (%)	Anticipated cost relative to incumbent in 2050 (summary, expert opinion survey and author group judgement)	Anticipated direction of cost relative to incumbent in 2050 (summary, expert opinion survey and author group judgement)
Polymers	1,440	2,040	-30%	Likely to be cheaper	Downward
Methanol	510	400	+30%	Insufficient consensus	Downward
Methane	1,740	360	+380%	Likely to be more expensive	Downward
Fischer-Tropsch fuels	4,160	1,200	+250%	Likely to be more expensive	Downward
Dimethyl ether	2,740	660	+320%	Insufficient consensus	Downward
Microalgae	2,680	1,000	+170%	Likely to be more expensive	Insufficient consensus
Aggregates	21	18	+20%	Insufficient consensus	Downward
Cement curing	56	71	-20%	Likely to be cheaper	Downward
CO <sub>2</sub> -EOR	n.a.	n.a.	n.a.	Likely to be more expensive	Upward

Median cost estimates for products made with CO<sub>2</sub> utilization are derived from the backward-looking scoping review. References for the selling prices are set out in more detail in Supplementary Table 4. The costs and cost trends anticipated in 2050 are derived from a forward-looking expert opinion survey and from author group judgement.

**Enhanced weathering.** The use of terrestrial enhanced weathering on croplands could increase crop yields<sup>28</sup>. This yield enhancement is unlikely to originate directly from increases in soil carbon, but from nutrient uptake that is facilitated by pH effects<sup>57</sup>. However, under our broad definition, there may still be an as-yet-unquantified CO<sub>2</sub> utilization potential associated with the increase in net primary productivity.

Forestry techniques. In afforestation/reforestation, atmospheric CO<sub>2</sub> is removed via photosynthesis and the carbon is stored in standing forests. If used for sustainable forestry, a portion of that carbon enters production processes and, after minor energetic losses, becomes wood products. Both wood products and standing forests provide economic value, and can be seen as CO<sub>2</sub> utilization (standing forests provide ecosystem services, which are not quantified here). The utilization of CO<sub>2</sub> in wood products will occur in addition to the direct removal of CO<sub>2</sub> by forests under certain highly specific circumstances; sustainable harvesting can maintain carbon stocks in forests while providing a source of renewable biomass<sup>58,59</sup>.

We estimate that, of the volumes of CO<sub>2</sub> sequestered via afforestation/reforestation in 2050, between 0.07 and 0.5 Gt of the CO<sub>2</sub> utilized per year may flow into industrial roundwood products, at approximate breakeven costs of between -\$40 and \$10 per tonne of CO<sub>2</sub> utilized. An optimistic scenario might also consider the volumes of wood products that are sustainably harvested from existing forests and plantations. Yearly inflows of carbon used as wood products are estimated to be around 1.8 Gt CO<sub>2</sub> in 2050. Of these, 0.6 Gt CO<sub>2</sub> may arise from the portion of those flows that are industrial roundwood products sustainably harvested for use in the construction industry (Supplementary Materials); this leads to a top-end estimate of 1.1 Gt CO<sub>2</sub> utilized per year from afforestation/reforestation and sustainable forestry techniques.

Wood products have potential as long-term stores of carbonparticularly when used in long-lived buildings, the lifespans of which can be conservatively estimated at 80–100 years<sup>59</sup>. We estimate that around half of the carbon in the wood-product pool might continue to be stored beyond the usable life of the products (the non-decomposed fraction of the portion of total wood products that are presently committed to landfill (around 60%) is approximately 77%<sup>60</sup>). The remainder of the carbon in the wood-product pool will return to the atmosphere as a fraction (about  $0.5\,\mathrm{Gt}\,\mathrm{CO_2}\,\mathrm{yr}^{-1}$ ) of the  $5\,\mathrm{Gt}\,\mathrm{CO_2}\,\mathrm{yr}^{-1}$  land-use change flux that is depicted in Fig. 1.

Soil carbon sequestration and biochar. CO<sub>2</sub> in land management and biochar pathways can be considered to be utilized if it enhances economically valuable agricultural output. The CO2 taken up by land ultimately becomes either CO<sub>2</sub> utilized (with increased output) or CO<sub>2</sub> removed (stored in soils), but not both. We estimate that around 0.9 to 1.9 Gt CO<sub>2</sub> yr<sup>-1</sup> may be used by soil carbon sequestration techniques on croplands and grazing lands by 2050; approximate breakeven costs are estimated at between -\$90 and -\$20 per tonne of CO<sub>2</sub> utilized, owing to yield increases that are associated with increases in soil organic carbon stock. We tentatively estimate that approximately 0.2 to 1Gt CO<sub>2</sub> yr<sup>-1</sup> may be utilized via yield increases after the application of biochar on managed lands, at approximate breakeven costs of between -\$70 and -\$60 per tonne of CO<sub>2</sub> utilized. These estimates are based on currently reported yield increases (of 0.9% to 2% associated with soil carbon sequestration techniques 61,62 and 10% associated with biochar 63) from sparse literature, using crop production as a proxy for net primary productivity. Impacts on yield are likely to be highly variable—for example, according to climatic zone<sup>64</sup>. Crop productivity increases are important not only for economic returns for operators but also for land-use requirements. For instance, if the application of biochar led to an increase in tropical biomass yields of 25%, the associated reduction in land requirements would equate to 185 million hectares, and would result in a cumulative net emission benefit from those increased yields of 180 Gt CO<sub>2</sub> to 2100<sup>65</sup>.

Table 2 presents breakeven cost ranges and estimated volumes of CO<sub>2</sub> utilized or removed per year in 2050.

# Techno-economic barriers to scaling

There are numerous challenges in scaling CO<sub>2</sub> utilization. Here we consider issues related to cost, technology and energy. Although market penetration can be facilitated by cost-competitiveness, there is no certainty that the cheapest CO<sub>2</sub> utilization pathways will scale up. Geographical, financing, political and societal considerations are briefly addressed in the Supplementary Materials; however, further investigation of these issues is warranted, particularly with regards to the UN Sustainable Development Goals.

# Cost and performance differentials

The breakeven cost per tonne of CO<sub>2</sub> is one way to assess the economics of utilization. The impact of CO<sub>2</sub> utilization on the price and value-add proposition of the end product is also important, particularly for CO<sub>2</sub> utilization processes in which the final price differential is immaterial but small differences in key properties may be important. Prices for a fuel product made using CO<sub>2</sub> currently exceed market prices considerably (Table 3).

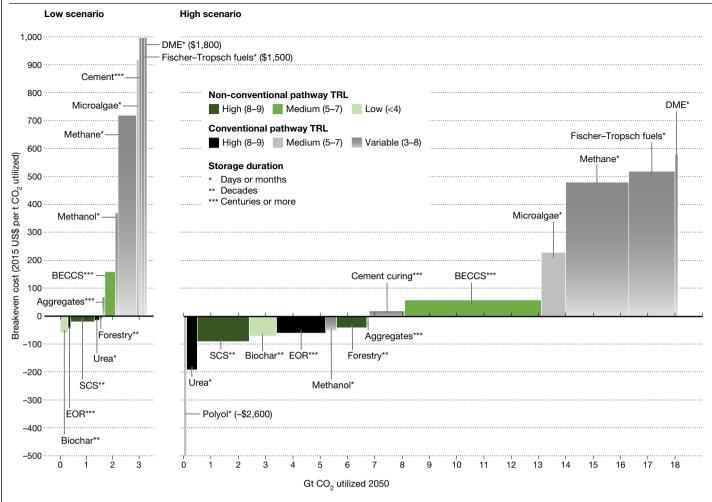


Fig. 3 | Estimated  $CO_2$  utilization potential and breakeven cost of different sub-pathways in low and high scenarios. The breakeven cost is the incentive, measured in 2015 US\$ per tonne of  $CO_2$ , that is required to make the pathway economic. Negative breakeven costs indicate that the pathway is already profitable, without any incentive to utilize  $CO_2$  (such as a tax on  $CO_2$  emissions in cases in which utilization avoids emissions, or a subsidy for  $CO_2$  removed from the atmosphere in the case in which utilization removes  $CO_2$ ). Utilization

estimates are based on 2050 projections. Many technologies are in the very early stages of development, and cost optimization via research and development could substantially change these estimates. Colour shadings reflect the TRLs of the pathways, which again vary markedly within each pathway. Asterisks denote the storage duration offered by each pathway: days or months (\*) decades (\*\*) or centuries or more (\*\*\*). See Supplementary Materials for further details.

Many of the other pathways—in particular those involving products in construction and plastics—have economics that are driven not only by price but also by the performance characteristics of the end product. There may be trade-offs between product quality and mitigation value, or synergies between the two.

Because they are based on a backward-looking scoping review, our cost estimates for conventional pathways do not capture current unpublished innovations and advances in the industrial arena. Our expert opinion survey, which included sources from both academia and industry, reflected great uncertainty about future costs. Industry participants expressed confidence that costs in pathways that are already economic (such as  $CO_2$  cement curing and polyols) would continue to decrease, relative to incumbent product costs.

# **Energy requirements**

Some  $CO_2$  utilization pathways involve chemical transformations that require the input of substantial amounts of energy (Supplementary Fig. 2). Some require energy to increase  $CO_2$  concentrations from 0.04% towards  $100\%^{66}$ . Life-cycle emissions and costs depend upon the source of the energy used. Land-based natural processes use solar energy, harnessed by photosynthesis, to transform  $CO_2$  and water into carbohydrates. Although photosynthesis is an inefficient process

(the average efficiency is around 0.2% globally<sup>67</sup>) biological pathways are not necessarily more expensive. In industrial processes, hydrogen often serves as feedstock. At present, 'brown' hydrogen is primarily—and most cheaply—generated by reforming methane<sup>68</sup>, which has associated  $CO_2$  emissions. In the production of 'blue' hydrogen, these emissions are captured and stored. Production of 'green' hydrogen—by the electrolysis of water— has real potential, and the ultimate choice of technology for the generation of hydrogen will depend on the rates of cost reduction<sup>69</sup>, among other factors.

# The outlook for CO<sub>2</sub> utilization

Our high-end and low-end scale and cost estimates in Table 2 are drawn as cost curves in low and high scenarios in Fig. 3. These curves are constructed using currently available (and often sparse) data in the peer-reviewed literature, or—where data are not available—using approximations, and should be considered as a speculative first pass at envisioning future scenarios. The curves should not be interpreted as comprehensive assessments of costs, they do not represent *n*th-of-a-kind costs, and they are incompatible with other sequestration or abatement cost curves. The limitations of cost curves—particularly with regards to exogenous costs such as establishment costs—have

been previously described<sup>70</sup>, and they remain relevant here. An important caveat is that individual potentials cannot be arbitrarily summed: some access the same demand, for instance for transport, which may or may not be filled by a process that utilizes CO<sub>2</sub>. For instance, the putative success of CO<sub>2</sub>-fuels may reduce the demand for oil, thus also reducing the potential of CO<sub>2</sub>-EOR. Furthermore, land availability means that choosing one land-based pathway (for example, BECCS) might preclude the application of another at scale (for example, biochar).

Notwithstanding the many caveats, the potential scale of utilization could be considerable. Much of this potential CO<sub>2</sub> utilizationnotably in 'closed' and 'open' pathways—may be economically viable without substantial shifts in prices. The specific assumptions of the low scenario, which do not account for potential overlaps in utilization volumes between pathways, imply an upper bound of over 1.5 Gt CO<sub>2</sub> yr<sup>-1</sup> at well under \$100 per tonne of CO<sub>2</sub> utilized. For policymakers that are interested in climate change, these figures demonstrate the theoretical potential for correctly designed policies to incentivize the displacement of fossil fuels or the removal of  $CO_2$  from the atmosphere.

Figure 3 also highlights some of the economic and technological challenges that are faced by these pathways. The cycling pathways (other than the production of urea and polyols) must compete with lower-cost incumbents. The four closed pathways, except for CO<sub>2</sub>-EOR, are mainly at low technology readiness levels (TRLs). Open pathways, although both theoretically profitable and implementable, often incur additional operating costs-such as implementation, transaction, institutional, and monitoring costs-which can be high71.

Each of the potentially large-scale, low-cost pathways also face challenges as mitigation strategies. CO2-EOR utilizes and, with correct policy, stores CO<sub>2</sub> at scale, but may not yield any net climate benefit and may even be detrimental. BECCS has a range of well-articulated risks, including considerable increases in emissions as a result of land-use change<sup>72</sup>. Land management, biochar and forestry offer only shorterterm storage, face saturation, and risk large-scale flows of CO<sub>2</sub> back to the atmosphere<sup>23</sup>. The chemicals pathways may reduce net emissions by displacing fossil fuel use, but will not contribute to net removal unless they are paired with direct air capture in a net zero world. Building materials face a challenging route to market penetration owing to regulatory barriers, which may take decades to surmount. In general, low TRLs will also challenge the ability of pathways to scale rapidly enough and within the desired timeframe for mitigation<sup>5</sup>. The uncertainty in future outcomes is relatively large, and very few industries globally involve over 1 Gt yr<sup>-1</sup> of material flows.

The net climate impact of the CO<sub>2</sub> utilization pathways will, in many cases, depend upon the emissions intensity from the prevailing processes<sup>73</sup>. For instance, CO<sub>2</sub>-EOR might currently contribute to an overall reduction in atmospheric CO<sub>2</sub>, compared to business-as-usual<sup>49</sup>. As decarbonization proceeds, however, the climate benefit of CO<sub>2</sub>-EOR is reduced. At some point before full decarbonization, EOR without direct air capture will result in a net increase in CO<sub>2</sub> emissions<sup>74</sup>. Conversely, in an economy with high supply-chain emissions, the climate benefit from BECCS is low<sup>72</sup>. In a decarbonized world, those supply-chain emissions will be close to zero and so the climate benefit from BECCS will be amplified.

Each of the utilization pathways described here should be seen as a part of the cascade of mitigation options that are available. For instance, using recycled organic matter to reduce fertilizer use and its associated emissions is a priority, followed by the more efficient use of fertilizer<sup>75</sup>, followed by increasing urea yields to reduce total emissions (via more efficient use of NH<sub>3</sub>)<sup>30</sup>. Eventually, fertilizers derived from fossilfuel-free ammonia for should be used to supplement fertilizers derived from organic materials. Similarly, a robust finding in the literature on integrated-assessment modelling is that the electricity sector should be decarbonized first, which then facilitates decarbonization in other, more difficult sectors<sup>77</sup>. In terms of the climate impact per kWh of electricity use, available renewable electricity is more efficiently directed towards e-mobility and heat pumps rather than towards hydrogen-based CCU technologies in the chemical industry<sup>73</sup>.

# Future priorities for CO<sub>2</sub> utilization

Given the slow nature of the innovation process and the urgency of the climate problem, priority should be given to the most promising and least-developed options so that early and effective adoption of a portfolio of techniques can be achieved. For the pathways with apparently negative cost (that is, those that should be profitable in the absence of a theoretical CO<sub>2</sub> subsidy), the challenge-particularly for the open pathways—is to identify and overcome the other barriers to adoption.

An important caveat for policymakers and practitioners is that scaling up CO<sub>2</sub> utilization will not necessarily be beneficial for climate stability; policy should not aim to support utilization per se, but should instead seek to incentivize genuine emission reductions and removals on a life-cycle basis, and thus provide incentives for the deployment of CO<sub>2</sub> utilization that is climate-beneficial.

#### Conventional utilization pathways

The emissions-reduction potentials of the three cycling pathways would be facilitated by declines in the costs of CO<sub>2</sub> capture. New sorbents could reduce the cost of energy-intensive separation of CO<sub>2</sub> from flue gases and industrial streams  $^{40,78}$ . In the longer term, cheaper direct air capture (based on clean energy) would support the scale-up of these pathways<sup>79</sup>. The cost of DACCS has recently been assessed to be between \$600 and \$1,000 per tonne of CO<sub>2</sub> for the first-of-a-kind plant, with nth-of-a-kind costs potentially of the order of \$200 per tonne of  $CO_2^{79}$ .

Research into materials and catalysts for CO<sub>2</sub> reduction could enable the efficient transformation of CO<sub>2</sub> into a broader range of products at a lower cost<sup>78</sup>. This includes the development of catalysts for the efficient production of syngas via dry reforming of methane with CO<sub>2</sub>; efficient photo/electrocatalysts to release hydrogen from water; photo/ electrocatalysts that can reduce CO<sub>2</sub>; or new high-temperature, reversibly reducible metal oxides<sup>78</sup> to produce syngas using concentrated sunlight. New membrane materials that can separate miscible liquidsfor example, methanol and water—will also be important<sup>80</sup>. Catalytic processes can be optimized to increase CO<sub>2</sub> emission reductions or to reduce energy consumption<sup>81</sup>. One important research challenge is to produce materials with the highest material property profiles. in particular temperature stability and wider operating or processing temperature windows. Rigorous, realistic techno-economic analyses of these scientific advances could determine their contribution to valuable cost reductions.

Given the rapid rate at which human societies are urbanizing<sup>82</sup>, there is an urgent one-time opportunity to deploy new building materials including wood, as discussed below—that utilize and store CO<sub>2</sub> and displace emissions-intensive Portland cement. In this area, as in others, progress would be aided by techno-economic analyses and life-cycle analyses with clearer system boundaries, counterfactuals, and accounting for co-products<sup>83</sup>, and integrated modelling frameworks that can co-assess changes in background systems<sup>84</sup>.

# Non-conventional utilization pathways

Figures 1 and 3 suggest that land-based biological processes offer a large opportunity to utilize, remove and store more CO<sub>2</sub>. Progress here is partly dependent upon field-based trials to improve understanding of the system-wide impacts of different pathways on plant yields and the impacts on water, food and water systems, and other resources. Such research might prioritize multiple-land-use approaches, such as agroforestry plantations; rice straw as biomass; low-displacement bioenergy strategies such crassulacean acid metabolism plants on marginal land; or nipa palm in mangroves. A better understanding of soil carbon dynamics and improved phenotypic and genotypic plant selection will also help<sup>85</sup>.

Biochar is currently at a low TRL and has associated uncertainties. However, if these can be overcome, its position low on the cost curve in both low and high scenarios suggests that this pathway may have considerable potential. A major challenge is to improve variations in yield effects, which are likely to hinder the economic decision made by farmers to apply biochar<sup>86</sup>, and to find ways to secure potential revenue streams.

Increased forestation, where land availability and biodiversity constraints allow, and the greater use of wood products in buildings are strategies that appear to be worth pursuing. Although our estimates consider the scale-up of existing industrial roundwood use via afforestation and reforestation, new wood-based products such as cross-laminated timber and acetylated wood  $^{87}$ —which are aimed at new markets—also have potential. Specification, quality and safety measures for these products are approaching comparability to many concrete structures  $^{88}$ , and current manufacturing scale-up suggests that this may be a market with strong growth prospects.

# **Cross-cutting efforts**

Broad policy and regulatory changes that may support the appropriate scale-up of CO<sub>2</sub> utilization include creating carbon prices of around \$40 to \$80 per tonne of CO<sub>2</sub>-increasing over time-to penalize CO<sub>2</sub> emissions <sup>89</sup> and to incentivize verifiable CO<sub>2</sub> emissions reductions and removals from the atmosphere. We do not advocate a direct subsidy for utilization. Instead, incentives for CO<sub>2</sub> removals and reductions (or penalties for emissions) are justified, and these will support CO<sub>2</sub> utilization in cases in which it is beneficial for the climate. For instance, our analysis suggests that closed pathways with scalability—such as BECCS and building materials—would be sensitive to a subsidy for CO<sub>2</sub> removals. Changes to standards, mandates, procurement policies and research and development support, in order to close gaps in knowledge across a portfolio of pathways 90, are also desirable. Financing and managing the emergence of a globally important new set of CO<sub>2</sub> utilization industries will probably require clear direction and industrial support from government. An enabling 'net zero' legislative regime-such as that in place in Sweden and the UK and proposed in New Zealand-can provide clarity about the necessary scale of industries that reduce and remove CO<sub>2</sub>, including the pathways examined here.

Collaboration between scholars, public officials and business leaders to ensure accurate comparisons between different alternatives—including the direct comparison of CCU, CDR and CCS pathways—could facilitate the blending of advantageous features of the ten pathways described here, the exploration of pathways not addressed here, and the identification of novel  $\mathrm{CO}_2$  utilization pathways to accelerate emissions reductions and removals.

 $CO_2$  utilization is not an end in itself, and these pathways solely or even collectively will not provide a key solution to climate change. Nevertheless, there is a substantial societal value in continued efforts to determine what will and will not work, in what contexts the climate will or will not benefit from  $CO_2$  utilization, and how expensive it will be.

- Dealing with Carbon Dioxide at Scale (The Royal Society and National Academy of Sciences, 2017).
- von der Assen, N. & Bardow, A. Life cycle assessment of polyols for polyurethane production using CO<sub>2</sub> as feedstock: insights from an industrial case study. Green Chem. 16, 3272–3280 (2014).
- Ampelli, C., Perathoner, S. & Centi, G. CO<sub>2</sub> utilization: an enabling element to move to a resource- and energy-efficient chemical and fuel production. *Philos. Trans. R. Soc. Lond.* A 373, 20140177 (2015).
- 4. The Potential and Limitations of Using Carbon Dioxide (The Royal Society, 2017).
- Mac Dowell, N., Fennell, P. S., Shah, N. & Maitland, G. C. The role of CO2 capture and utilization in mitigating climate change. Nat. Clim. Change 7, 243–249 (2017).
   This paper assesses the potential for CO<sub>2</sub>-derived fuels and chemicals to be a fraction of that possible via CO<sub>2</sub>-EOR.
- 6. IPCC Special Report: Carbon Dioxide Capture and Storage (eds Metz, B., Davidson, O. R., De Coninck, H., Loos, M. & Meyer, L. A.) (Cambridge Univ. Press, 2005).
  This IPCC report provides an overview of the technology and expected costs of carbon capture and sequestration, and provides a key definition of CO<sub>2</sub> utilization.

- Aresta, M., Dibenedetto, A. & Angelini, A. Catalysis for the valorization of exhaust carbon: from CO<sub>2</sub> to chemicals, materials, and fuels. Technological use of CO<sub>2</sub>. Chem. Rev. 114, 1709–1742 (2014).
- Quadrelli, E. A., Centi, G., Duplan, J. L. & Perathoner, S. Carbon dioxide recycling: emerging large-scale technologies with industrial potential. *ChemSusChem* 4, 1194–1215 (2011)
- Mikkelsen, M., Jorgensen, M. & Krebs, F. C. The teraton challenge. A review of fixation and transformation of carbon dioxide. Energy Environ. Sci. 3, 43–81 (2010).
- Markewitz, P. et al. Worldwide innovations in the development of carbon capture technologies and the utilization of CO<sub>2</sub>. Energy Environ. Sci. 5, 7281–7305 (2012).
- Bushuyev, O. S. et al. What should we make with CO2 and how can we make it? Joule 2, 825–832 (2018)
- Majumdar, A. & Deutch, J. Research opportunities for CO2 utilization and negative emissions at the gigatonne scale. *Joule* 2, 805–809 (2018).
  - This high-level commentary proposes, using industrial methods, harnessing of the natural biological cycle and a systems approach for industrial CO<sub>2</sub> utilization at scale.

    Bennett, S. J., Schroeder, D. J. & McCov, S. T. Towards a framework for discussing and
- assessing CO<sub>2</sub> utilisation in a climate context. *Energy Procedia* **63**, 7976–7992 (2014).
- Harper, A. B. et al. Land-use emissions play a critical role in land-based mitigation for Paris climate targets. Nat. Commun. 9, 2938 (2018).
- 15. Le Quéré, C. et al. Global carbon budget 2018. Earth Syst. Sci. Data 10, 2141–2194 (2018).
- IPCC Climate Change 2014: Mitigation of Climate Change (eds. Edenhofer, O. et al.) (Cambridge Univ. Press, 2014).
- 17. Minasny, B. et al. Soil carbon 4 per mille. Geoderma 292, 59-86 (2017).
- Smith, P. et al. Biophysical and economic limits to negative CO<sub>2</sub> emissions. Nat. Clim. Change 6, 42–50 (2016).

# This paper quantifies potential global impacts of various negative emissions technologies in the context of biophysical resource constraints.

- Lauk, C., Haberl, H., Erb, K.-H., Gingrich, S. & Krausmann, F. Global socioeconomic carbon stocks in long-lived products 1900–2008. Environ. Res. Lett. 7, 034023 (2012).
- Xi, F. et al. Substantial global carbon uptake by cement carbonation. Nat. Geosci. 9 880–883 (2016).
- Maries, A., Tyrer, M. & Provis, J. L. Sequestration of CO<sub>2</sub> emissions from cement manufacture. In Proc. 37th Cement and Concrete Science Conference (eds Bai, Y. et al.) (Institute of Materials, Minerals and Mining, 2017).
- Alcalde, J. et al. Estimating geological CO<sub>2</sub> storage security to deliver on climate mitigation. Nat. Commun. 9, 2201 (2018).
- Baccini, A. et al. Tropical forests are a net carbon source based on aboveground measurements of gain and loss. Science 358, 230–234 (2017).
- Allen, C. D. et al. A global overview of drought and heat-induced tree mortality reveals emerging climate change risks for forests. For. Ecol. Manage. 259, 660–684 (2010).
- Scott, V., Haszeldine, R. S., Tett, S. F. B. & Oschlies, A. Fossil fuels in a trillion tonne world. Nat. Clim. Change 5, 419 (2015).
- Gardiner, S. M. A perfect moral storm: climate change, intergenerational ethics and the problem of moral corruption. *Environ. Values* 15, 397–413 (2006).
- Naims, H. Economics of carbon dioxide capture and utilization—a supply and demand perspective. Environ. Sci. Pollut. Res. 23, 22226–22241 (2016).
   This paper analyses CO<sub>2</sub> supply and demand scenarios to conclude that the business
  - In is paper analyses CO<sub>2</sub> supply and demand scenarios to conclude that the business case for CO<sub>2</sub> utilization is technology-specific.
- Fuss, S. et al. Negative emissions—Part 2: Costs, potentials and side effects. Environ. Res. Lett. 13, 063002 (2018).
  - This paper estimates—through a large scoping review—that afforestation and reforestation, BECCS, biochar, enhanced weathering, DACCS and soil carbon sequestration all have multi-gigatonne sequestration potentials in 2050, and that costs vary widely.
- Otto, A., Grube, T., Schiebahn, S. & Stolten, D. Closing the loop: captured CO<sub>2</sub> as a feedstock in the chemical industry. Energy Environ. Sci. 8, 3283–3297 (2015).
- Pérez-Fortes, M., Bocin-Dumitriu, A. & Tzimas, E. CO<sub>2</sub> utilization pathways: Technoeconomic assessment and market opportunities. Energy Procedia 63, 7968–7975 (2014).
- Langanke, J. et al. Carbon dioxide (CO<sub>2</sub>) as sustainable feedstock for polyurethane production. Green Chem. 16, 1865–1870 (2014).
- Shih, C. F., Zhang, T., Li, J. & Bai, C. Powering the future with liquid sunshine. Joule, 2, 1925–1949 (2018).
- Jarvis, S. M. & Samsatli, S. Technologies and infrastructures underpinning future CO<sub>2</sub> value chains: A comprehensive review and comparative analysis. *Renew. Sustain. Energy Rev.* 85, 46–68 (2018).
- Myhre, G. et al. In Climate Change 2013: The Physical Science Basis (eds Stocker, T. F. et al.) 659–740 (IPCC, Cambridge Univ. Press, 2013).
- Luo, J., Ledgard, S. & Lindsey, S. Nitrous oxide emissions from application of urea on New Zealand pasture. N. Z. J. Agric. Res. 50, 1–11 (2007).
- Geyer, R., Jambeck, J. R. & Law, K. L. Production, use, and fate of all plastics ever made Sci. Adv. 3, e1700782 (2017).
- Jiang, Z., Xiao, T., Kuznetsov, V. L. & Edwards, P. P. Turning carbon dioxide into fuel. *Philos. Trans. A* 368, 3343–3364 (2010).
- Olah, G. A. Beyond oil and gas: the methanol economy. Angew. Chem. Int. Ed. 44, 2636–2639 (2005).
- National Academies of Sciences, Engineering, and Medicine. Commercial Aircraft Propulsion and Energy Systems Research: Reducing Global Carbon Emissions (National Academies Press, 2016).
- Secretary of Energy Advisory Board. Letter Report: Task Force on RD&D Strategy for CO<sub>2</sub>
   Utilization and/or Negative Emissions at the Gigatonne Scale. (US Department of Energy, 2016).
- 41. De Luna, P. et al. What would it take for renewably powered electrosynthesis to displace petrochemical processes? Science 364, eaav3506 (2019). This paper reviews the potential for and costs of using renewable energy for electrochemical conversion of concentrated CO<sub>2</sub> into formic acid, carbon monoxide, ethylene and ethanol, and compares biocatalytic and Fischer-Tropsch routes to long-chain chemical production

- Dimitriou, I. et al. Carbon dioxide utilisation for production of transport fuels: process and economic analysis. Energy Environ. Sci. 8, 1775-1789 (2015).
- 43. Laurens, L. M. L. State of Technology Review Algae Bioenergy (IEA Bioenergy, 2017).
- 44. Brynolf, S., Taljegard, M., Grahn, M. & Hansson, J. Electrofuels for the transport sector: A review of production costs. Renew. Sustain. Energy Rev. 81, 1887-1907 (2018).
- Williams, P. J. B. & Laurens, L. M. Microalgae as biodiesel & biomass feedstocks: review & analysis of the biochemistry, energetics & economics. Energy Environ. Sci. 3, 554-590
- 46. Mahoutian, M. & Shao, Y. Production of cement-free construction blocks from industry wastes, J. Clean, Prod. 137, 1339-1346 (2016).
- Provis, J. L. & Bernal, S. A. J. Geopolymers and related alkali-activated materials. Annu. Rev. Mater. Res. 44, 299-327 (2014).
- 48. Dai, Z. et al. CO<sub>2</sub> accounting and risk analysis for CO<sub>2</sub> sequestration at enhanced oil recovery sites. Environ. Sci. Technol. 50, 7546-7554 (2016).
- Heidug, W. et al. Storing CO<sub>2</sub> through enhanced oil recovery; combining EOR with CO<sub>2</sub> 49 storage (EOR+) for profit. (International Energy Agency, 2015).
- Stewart, R. J. & Haszeldine, R. S. Can producing oil store carbon? Greenhouse gas 50. footprint of CO<sub>2</sub>EOR, offshore North Sea, Environ, Sci. Technol, 49, 5788-5795 (2015).
- Godec, M. L. Global Technology Roadmap for CCS in Industry: Sectoral Assessment CO. 51. Enhanced Oil Recovery, (United Nations Industrial Development Organization, 2011).
- Griscom, B. W. et al. Natural climate solutions. *Proc. Natl Acad. Sci. USA* **114**, 11645–11650 52 (2017).
- 53 Smith, P. Soil carbon sequestration and biochar as negative emission technologies. Glob. Change Biol. 22, 1315-1324 (2016).
- 54. Minx, J. C. et al. Negative emissions—Part 1: Research landscape and synthesis. *Environ*. Res. Lett. 13, 063001 (2018).
- 55 Slade, R., Bauen, A. & Gross, R. Global bioenergy resources. Nat. Clim. Change 4, 99
- 56. Vaughan, N. E. et al. Evaluating the use of biomass energy with carbon capture and storage in low emission scenarios. Environ. Res. Lett. 13, 044014 (2018).
- Beerling, D. J. et al. Farming with crops and rocks to address global climate, food and soil security. Nat. Plants 4, 138-147 (2018).
- Pingoud, K., Ekholm, T., Sievänen, R., Huuskonen, S. & Hynynen, J. Trade-offs between forest carbon stocks and harvests in a steady state - a multi-criteria analysis. J. Environ. Manage. 210, 96-103 (2018).
- Lippke, B. et al. Life cycle impacts of forest management and wood utilization on carbon mitigation: knowns and unknowns. Carbon Manage. 2, 303-333 (2011).
- 60. FAOSTAT (Food and Agricultural Organization of the United Nations, accessed 10 May 2018); http://fao.org/faostat/en/#data
- Lal, R. Enhancing crop yields in the developing countries through restoration of the soil 61 organic carbon pool in agricultural lands. Land Degrad, Dev. 17, 197-209 (2006).
- 62. Soussana, J.-F. et al. Matching policy and science: Rationale for the '4 per 1000-soils for food security and climate' initiative. Soil Tillage Res. 188, 3-15 (2019).
- leffery S. Verheijen, F.G. Van Der Velde, M. & Bastos, A. C. A quantitative review of the 63 effects of biochar application to soils on crop productivity using meta-analysis. Agric. Ecosyst, Environ, 144, 175-187 (2011).
- 64 Jeffery, S. et al. Biochar boosts tropical but not temperate crop yields. Environ. Res. Lett. 12. (2017).
- Werner, C., Schmidt, H. P., Gerten, D., Lucht, W. & Kammann, C. Biogeochemical potential of biomass pyrolysis systems for limiting global warming to 1.5 °C. Environ. Res. Lett. 13, (2018)
- Darton, R. & Yang, A. Removing carbon dioxide from the atmosphere assessing the technologies. Chem. Eng. Trans. 69, 91-96 (2018).
- Barber, J. Photosynthetic energy conversion: natural and artificial. Chem. Soc. Rev. 38, 185-196 (2009).
- Izquierdo, U. et al. Hydrogen production from methane and natural gas steam reforming in conventional and microreactor reaction systems. Int. J. Hydrogen Energy 37,
- 69. Kuckshinrichs, W., Ketelaer, T. & Koj, J. C. Economic analysis of improved alkaline water electrolysis. Front. Energy Res. 5, 1 (2017).
- Kesicki, F. & Strachan, N. Marginal abatement cost (MAC) curves: confronting theory and practice. Environ. Sci. Policy 14, 1195-1204 (2011).
- 71 Viana, V. M., Grieg-Gran, M., Della Mea, R. & Ribenboim, G. The Costs of REDD: Lessons From Amazonas (International Institute for Environment and Development, 2009).
- Faiardy, M. & Mac Dowell, N. Can BECCS deliver sustainable and resource efficient negative emissions? Energy Environ, Sci. 10, 1389-1426 (2017).
- 73. Kätelhön, A., Meys, R., Deutz, S., Suh, S. & Bardow, A. Climate change mitigation potential of carbon capture and utilization in the chemical industry. Proc. Natl Acad. Sci. USA 116, 11187-11194 (2019).
- Jaramillo, P., Griffin, W. M. & McCoy, S. T. Life cycle inventory of  $\mathrm{CO}_2$  in an enhanced oil recovery system. Environ. Sci. Technol. 43, 8027-8032 (2009).
- Gerber, J. S. et al. Spatially explicit estimates of  $N_2O$  emissions from croplands suggest climate mitigation opportunities from improved fertilizer management. Glob. Change Biol. 22, 3383-3394 (2016).
- Chen, J. G. et al. Beyond fossil fuel-driven nitrogen transformations. Science 360, 76. eaar6611 (2018).
- Luderer, G. et al. Residual fossil  $CO_2$  emissions in 1.5–2°C pathways. Nat. Clim. Change 8, 77. 626-633 (2018).
- Senftle, T. P. & Carter, E. A. The holy grail: chemistry enabling an economically viable  ${\rm CO_2}$ capture, utilization, and storage strategy. Acc. Chem. Res. 50, 472-475 (2017).
- Keith, D. W., Holmes, G., St., Angelo, D. & Heidel, K. A process for capturing  $\mathrm{CO}_2$  from the atmosphere, Joule 2, 1573-1594 (2018).
- 80. Mahmood, A., Bano, S., Kim, S.-G. & Lee, K.-H. Water-methanol separation characteristics of annealed SA/PVA complex membranes. J. Membr. Sci. 415-416, 360-367 (2012).
- Xiao, T. et al. The Catalyst Selectivity Index (CSI): a framework and metric to assess the impact of catalyst efficiency enhancements upon energy and CO2 footprints. Top. Catal. **58**, 682-695 (2015).

- Seto, K. C., Güneralp, B. & Hutyra, L. R. Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools. Proc. Natl Acad. Sci. USA 109, 16083-16088 (2012).
- Zimmermann, A. et al. Techno-Economic Assessment & Life-Cycle Assessment Guidelines for CO<sub>2</sub> Utilization (Global CO<sub>2</sub> Initiative, 2018).
- Arvesen, A., Luderer, G., Pehl, M., Bodirsky, B. L. & Hertwich, E. G. Deriving life cycle assessment coefficients for application in integrated assessment modelling. Environ. Model, Softw. 99, 111-125 (2018).
- Scharlemann, J. P. W., Tanner, E. V. J., Hiederer, R. & Kapos, V. Global soil carbon: understanding and managing the largest terrestrial carbon pool. Carbon Manage. 5,
- Dickinson, D. et al. Cost-benefit analysis of using biochar to improve cereals agriculture. Glob. Change Biol. Bioenergy 7, 850-864 (2015).
- Song, J. et al. Processing bulk natural wood into a high-performance structural material. Nature **554**, 224-228 (2018).
- Ramage, M. H. et al. The wood from the trees: The use of timber in construction. Renew. Sustain, Energy Rev. 68, 333-359 (2017).
- High-Level Commission on Carbon Prices Report of the High-Level Commission on 89. Carbon Prices (World Bank, 2017).
- Hepburn, C., Pless, J. & Popp, D. Encouraging innovation that protects environmental systems; five policy proposals, Rev. Environ, Econ. Policy (2018).
- Muntean, M. et al. Fossil CO<sub>2</sub> Emissions of all World Countries—2018 Report. EUR 29433 EN, JRC113738 (Publications Office of the European Union, 2018).
- 92. Sundquist, E. & Visser, K. The geologic history of the carbon cycle. Treatise Geochem. 8, 682 (2003).
- 93. Blunden, J., Derek, S. & Hartfield, G. State of the Climate in 2017. Bull. Amer. Meteor. Soc. 99, Si-S310 (2018).
- Cuéllar-Franca, R. M. & Azapagic, A. Carbon capture, storage and utilisation technologies: a critical analysis and comparison of their life cycle environmental impacts. J. CO<sub>2</sub> Utilization 9, 82-102 (2015).

#### This paper compares the environmental impacts of CO<sub>2</sub> utilization and CCS technologies by reviewing the literature of life cycle assessment studies

- Sathre, R. & O'Connor, J. Meta-analysis of greenhouse gas displacement factors of wood product substitution. Environ. Sci. Policy 13, 104-114 (2010).
- van der Giesen, C., Kleijn, R. & Kramer, G. J. Energy and climate impacts of producing synthetic hydrocarbon fuels from CO<sub>2</sub>. Environ. Sci. Technol. 48, 7111-7121 (2014).
- Sternberg, A., Jens, C. M. & Bardow, A. Life cycle assessment of CO<sub>2</sub>-based C1-chemicals. 97. Green Chem. 19, 2244-2259 (2017).
- Abanades, J. C., Rubin, E. S., Mazzotti, M. & Herzog, H. J. On the climate change mitigation potential of CO<sub>2</sub> conversion to fuels. Energy Environ. Sci. 10, 2491-2499 (2017).
- Sternberg, A. & Bardow, A. Life cycle assessment of power-to-gas: syngas vs methane. ACS Sustain. Chem. Eng. 4, 4156-4165 (2016).

Acknowledgements We thank the participants at the 2017 Sackler Forum of the UK Royal Society and the US National Academy of Sciences for input and critique on an earlier related discussion paper. We thank T. Chen, A. Cheng, Y. Lu, T. Ooms, R. Rafaty, V. Schreiber and A. Stephens for research assistance; and J. Adams, R. Aines, M. Allen, D. Beerling, P. Carey, I. Dairanieh, R. Darton, M. Davidson, R. Davis, B. David, N. DeCristofaro, N. Deich, P. Edwards, J. Fargione, J. Friedmann, S. Gardiner, A. Gault, C. Godfray, G. Henderson, K. Hortmann, S. Hovorka, G. Hutchings, D. Keith, J. King, T. Kruger, G. Lomax, M. Mason, S. McCoy, A. Mehta, H. Naims, T. Schuler, R. Sellens, N. Shah, P. Styring, J. Wilcox and E. Williams for their ideas and critique, although this should not be taken as implying their approval or agreement with anything in this paper. We thank participants at the 2018 CCS Forum in Italy, and participants at the 2019 Oxford Energy Colloquium. We thank J. Ditner for drawing the initial version of Fig. 1. This work was funded primarily by the Oxford Martin School, with other support from The Nature Conservancy, S.F. and J.C.M. have contributed to this work under the Project 'Strategic Scenario Analysis' (START) funded by the German Ministry of Research and Education (grant reference: 03EK3046B). The input of P.S. contributes to the Belmont Forum/FACCE-JPI DEVIL project (NE/MO21327/1) and the Natural Environment Research Council (NERC)-funded Soils-R-GGREAT project (NE/P019455/1) and the UKERC-funded Assess-BECCS project. The contribution of N.M.D. is funded by 'Region-specific optimisation of greenhouse gas removal' funded by NERC, under grant NE/P019900/1. The input of E.A.C. is funded by the US Air Force Office of Scientific Research, award number FA9550-14-1-0254.

Author contributions J.B. conceived of the paper. C.H. and E.A. conducted the analysis and drafted the paper, with extensive input from N.M.D., and critical input on estimates, methodology and drafting from J.B., E.A.C., S.F., J.C.M., P.S. and C.K.W.

Competing interests C.H. has funding from The Nature Conservancy and in the past has had funding from Shell. He is a Director of Vivid Economics, an economics consultancy firm. N.M.D. has funding from COSIA. Shell and Total, consults for BP, and has consulted in the past for Exxon, C.K.W. is a Director of Econic Technologies.

#### Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-

Correspondence and requests for materials should be addressed to E.A.

Peer review information Nature thanks Andrea Ramirez Ramirez, Keywan Riahi and the other, anonymous, reviewer(s) for their contribution to the peer review of this work

Reprints and permissions information is available at http://www.nature.com/reprints. Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© Springer Nature Limited 2019

# Anatomy and resilience of the global production ecosystem

https://doi.org/10.1038/s41586-019-1712-3

Received: 28 July 2018

Accepted: 23 September 2019

Published online: 6 November 2019

M. Nyström1\*, J.-B. Jouffray12, A. V. Norström1, B. Crona12, P. Søgaard Jørgensen12, S. R. Carpenter<sup>3</sup>, Ö. Bodin<sup>1</sup>, V. Galaz<sup>1,2</sup> & C. Folke<sup>1,2,4</sup>

Much of the Earth's biosphere has been appropriated for the production of harvestable biomass in the form of food, fuel and fibre. Here we show that the simplification and intensification of these systems and their growing connection to international markets has yielded a global production ecosystem that is homogenous, highly connected and characterized by weakened internal feedbacks. We argue that these features converge to yield high and predictable supplies of biomass in the short term, but create conditions for novel and pervasive risks to emerge and interact in the longer term. Steering the global production ecosystem towards a sustainable trajectory will require the redirection of finance, increased transparency and traceability in supply chains, and the participation of a multitude of players, including integrated 'keystone actors' such as multinational corporations.



The demand for harvestable biomass (food, fuel and fibre) by a growing, wealthier and increasingly urbanized global human population is placing relentless pressure on the Earth's ecosystems. To a large extent. this demand has been met by converting ecosystems into production ecosystems – ecosystems modified for the production of one or a few harvestable species<sup>1,2</sup>. Although these alterations occur at local scales, their cumulative effect is causing global transformation of the Earth's biosphere  $^{3,4}.$  Humans have already altered more than 75% of the world's terrestrial habitats<sup>5</sup>-nearly 40% of all productive land has been converted into agricultural areas<sup>6</sup> and two thirds of all boreal forests are under some form of management, mainly for wood production<sup>7</sup>. In the seas, around 90% of large industrial fisheries are either overexploited or fully exploited8, and a rapidly expanding aquaculture sector is occupying increasing areas of coastal and offshore space9.

As available productive land and abundant fish stocks become progressively scarce, the potential for further land conversion, land redistribution and exploitation of new wild stocks as options to meet projected global human demand is dwindling<sup>8,10</sup>. To increase efficiency, production ecosystems are intensified and simplified using human inputs such as fossil fuels, fertilizers, pesticides, antibiotics and technology<sup>2,6,11</sup>. In parallel, people, places, cultures and economies are increasingly linked across geographic locations and socioeconomic contexts<sup>12</sup>, making production ecosystems increasingly globally interconnected. Collectively, these changes are converting much of the biosphere into a GPE.

This new reality calls for approaches that recognize the biosphere system as a complex and integrated social-ecological system<sup>3,13,14</sup>. Within this context, resilience—the capacity of a system to persist with and adapt to change, but also transform away from unsustainable socialecological trajectories-has been suggested as a conceptual framework that could assist in developing paths towards sustainability<sup>15</sup>. Whereas the aggregated transformation of Earth's biomes is indisputable, its consequences for the dynamics and resilience of an expanding GPE remain poorly understood.

Here we describe the anatomy of the GPE through the lens of three key features underpinning resilience, namely connectivity, diversity and feedback<sup>16</sup>. We do this by considering a diverse set of socioeconomic and biophysical elements that have previously been studied separately. We discuss how this anatomy influences the resilience of the GPE and creates novel conditions for risks to emerge and interact. We conclude by highlighting three avenues that can foster innovation and encourage new partnerships to motivate transformation towards a more sustainable GPE.

# The anatomy of the GPE

The GPE is the result of three important and interacting trends: (1) the continued conversion of the Earth's biosphere into simplified production ecosystems, (2) the increased intensification and dependence of these production ecosystems on human inputs, and (3) their expanding connectivity through global markets. The GPE integrates multiple sectors, broadly referred to here as forestry, agriculture (crops and livestock) and fishery (wild capture and aquaculture) (Fig. 1). We recognize that some production ecosystems, such as subsistence fishing and farming or diversified agricultural landscapes, may be subject to

Stockholm Resilience Centre, Stockholm University, Stockholm, Sweden, 2Global Economic Dynamics and the Biosphere Academy Programme, Royal Swedish Academy of Sciences, Stockholm, Sweden. 3 Center for Limnology, University of Wisconsin-Madison, Madison, WI, USA. 4 Beijer Institute of Ecological Economics, Royal Swedish Academy of Sciences, Stockholm, Sweden, \*e-mail: magnus.nvstrom@su.se

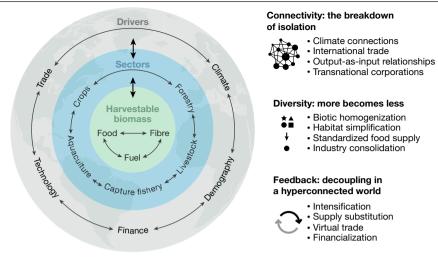


Fig. 1| The global production ecosystem. The GPE is characterized by tightly coupled relationships and reciprocal influence within and between harvestable biomass (green inner circle), multiple sectors (blue middle circle)

and a broad set of distal drivers (grey outer circle). To the right are the three lenses (connectivity, diversity and feedback) and their key features through which the anatomy of the GPE is described in this paper.

little human input or export-mediated connectivity from international trade. Nevertheless, they will be increasingly shaped by a broader set of global drivers, such as policies, technologies and economic changes<sup>2,17</sup>.

#### Connectivity: the breakdown of isolation

A distinctive feature of the present day is the way in which human activities increase connectivity. Although the drivers of this connectivity are not new (for example, trade, transport, technology and consumption), the speed and scale at which it occurs are unprecedented<sup>18</sup>.

Connectivity within the GPE is underpinned by long-distance biophysical and socioeconomic teleconnections 12,14. For example, irrigation and deforestation for agriculture in one location can redistribute global evapotranspiration, thereby changing rainfall patterns and affecting terrestrial production ecosystems in other regions<sup>19</sup>. Increased CO<sub>2</sub> emissions associated with deforestation<sup>20</sup> also affect aquaculture and wild-capture fisheries through increased seawater temperatures and ocean acidification<sup>21</sup>. Thus, land transformation in one part of the world can have substantial effects on production ecosystems at distant locations, within and across sectors.

At the same time, trade that was once constrained by limitations in transport capacities and lack of trade agreements is increasingly contributing to match global supply and demand<sup>22,23</sup>. International trade has undergone huge expansion in the past few decades<sup>24</sup>, and now accounts for 24% of all agricultural land25, 23% of the freshwater resources used for food production<sup>26</sup> and more than 35% of global seafood production<sup>8</sup>. The number of regional trade agreements in force has more than tripled<sup>27</sup> since 2000, and nearly all cropland areas brought into production from 1986 to 2009 were used to grow export crops<sup>28</sup>. As a consequence, production ecosystems have been further simplified and intensified to produce products destined for global markets<sup>28-31</sup>.

The growth of international trade has also increased direct and indirect connections between different production ecosystem sectors. For example, agricultural exports such as soybean and palm oil produced for the European Union, US and Chinese markets are a primary driver of deforestation across the tropics<sup>32</sup>. Sectors have also become intertwined through different output-as-input relationships. Increase in feed trade to satisfy global livestock production is occurring at an unprecedented rate<sup>33</sup>, and as the effects of intensification unfold, new connections are emerging. For instance, the aquaculture sector, which has traditionally relied heavily on capture fisheries as the main source for feed, is shifting towards agriculture for crop-based feed (for example, soy, rapeseed and maize) in response to declining fish catches<sup>34</sup>.

The interconnections between sectors are further amplified by the emergence of large transnational corporations that link production ecosystems globally through their subsidiaries<sup>35</sup>. These vertically and horizontally integrated 'keystone actors' <sup>35</sup> rely on connectivity for their own growth and represent a critical feature of the GPE by operating across sectors, markets and geographies to source, store, trade, process and distribute biomass. Such integration allows a few actors to dominate all segments of production, control the whole supply chain and have a disproportionate influence on decision-making<sup>36</sup>. Consolidation of large industrial actors has been recorded across many sectors, including forestry, seafood, livestock and agri-food industries <sup>37,38</sup>. There are concerns that such consolidation reinforces global homogenization of species (including genes, varieties and crops), practice and knowledge  $^{39,40}$ .

#### Diversity: more becomes less

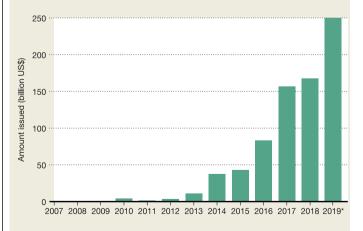
The purposeful selection of particular harvestable products and the collateral effects of these choices are driving biotic homogenization in both terrestrial and aquatic ecosystems 41,42. In many areas, boreal forests have been simplified as a consequence of intensive silviculture for timber production<sup>7</sup>, tropical forests have been replaced by spatially extensive monocultures (for example, soy and oil palm plantations)<sup>43</sup>, and native Mediterranean ecosystems have been simplified by exotic pine tree plantations<sup>44</sup>. In grasslands, moderate intensification has resulted in collateral biotic homogenization across microbial, plant and animal groups, both above and below ground<sup>45</sup>. In the Amazon, rainforest bacterial communities have become homogenized as a result of land conversion to cattle pasture<sup>46</sup> and in marine systems, rising seawater temperatures have led to the rapid homogenization of fish assemblages<sup>47</sup>.

Homogenization is also evident from a food production perspective. More than 80% of the global fish and shellfish aquaculture production is sourced from 30 species, of which grass carp, silver carp, cupped oysters, common carp and manila clam account for more than 30% by volume<sup>8</sup>. The pattern is even more striking for the global livestock sector, in which the production of pigs and chicken amount to 40% and 34%, respectively, of global meat production<sup>48</sup>. In agriculture, national portfolios of food supplies have seen increased crop species diversity, whereas globally they have become more homogeneous in composition, illustrating a shift towards a globally standardized food supply based on a few crop types such as maize, wheat, rice and barley<sup>49</sup>. Homogenization of crop production is further promoted by the

# Box 1

# Financialization of the biosphere

After decades of financial deregulation and innovation, the intensification of investments in natural assets has led to concerns about the role of an expanding global financial sector in shaping production ecosystems 139,140. Financialization—defined as the increasing importance of financial markets, motives, institutions and elites in the operation of the economy and its governing institutions<sup>141</sup>—has been suggested as a rapidly emerging and powerful decoupling mechanism that abstracts biomass from its physical form<sup>53</sup>. For example, new financial practices and instruments, such as securitization and complex commodity derivatives, have led to a situation where only 2% of commodity futures contracts end with delivery of the physical good<sup>93</sup>. In environmental conservation, financialization is seen as a new frontier for capital investment, in which the responsibility for global environmental outcome is increasingly shifted towards the incentivizing control of investment finance<sup>142</sup>. The green bond market, which is designed to simultaneously achieve financial returns and environmental benefits, has witnessed huge growth over the past five years and is predicted to reach US\$250 billion in 2019 (Box 1 Figure). Attention is also shifting towards the role of finance within the blue economy narrative 143. Similar to green bonds, the Seychelles government has announced the world's first blue bond, valued at US\$20 million to fund sustainable fisheries<sup>144</sup>. However, studies have warned that the growing interest among financial institutions in investments in the seafood sector may lead to adverse effects on small-scale fisheries through increased privatization and ocean grabbing<sup>145</sup>. The use of fishing quotas as collateral in loans by Icelandic banks, and the resulting debt for the industry when the banking system collapsed, provides a compelling example<sup>146</sup>. If financial actors were to become aware of how ecological risks translate into financial risks, entry points for sustainability considerations into financial decisions might emerge with strong incentives to implement better standards and redirect capital towards more sustainable practices<sup>113</sup>. Regardless, financialization of the biosphere should be increasingly recognized and studied as an intrinsic process shaping the GPE.



Box 1 Figure | Rise of the global green bond market. Data are from the Climate Bond Initiative (https://www.climatebonds.net). The 2019 data are forecast (indicated by an asterisk). Note that while the growth of the green bond market is unquestionable, it represents only a tiny fraction of the overall bond market, estimated to exceed US\$100 trillion.

recent rise of 'flex crops and commodities'. These are commodities that are suited for multiple uses that can be flexibly interchanged (for example, soy as food for humans, feed for animals, or biofuel; or trees for timber, pulp, ethanol, or carbon sequestration purposes). Such commodities provide flexibility for producers and investors to allocate products depending on which market has the highest demand—for instance, in the face of changes in policy regulations, market prices or technological advances<sup>50</sup>.

# Feedback: decoupling in a hyperconnected world

Paradoxically, increased connectivity within and among production ecosystems is weakening important feedback relationships within the GPE. First, there is broad evidence that intensification decouples production ecosystems from the natural processes needed to sustain desired production outcomes (that is, regulating and supporting ecosystem services)2. Instead, human inputs are increasingly used to mimic natural processes and responses in the system. Examples include substituting the natural breakdown and uptake of nutrients (that is, nutrient recycling) with fertilizers to enhance crop productivity, relying on artificial feed inputs to increase aquaculture yield, and replacing natural pest control with pesticides and herbicides to avoid yield losses. This can potentially undermine the capacity of production ecosystems to sustain desired biomass in the long run. For instance, agricultural intensification has been reported to cause soil erosion, declines in fertility, loss of natural pollinators, downstream damage to water resources and degradation of coastal ecosystems<sup>38,51</sup>.

Decoupling also emerges as the geographical distance between the location of biomass production and where it is consumed increases. Approximately one quarter of all food produced for human consumption is traded internationally<sup>24</sup>, and almost one billion people are consuming internationally traded products to cover their daily nutrition<sup>52</sup>. Estimates further suggest that 20% of global cropland is being allocated to the production of commodities that are consumed in another country<sup>31</sup>. This spatial decoupling, or 'distancing'53, allows industries to substitute supplies from different species or production ecosystems so that global consumers remain relatively unaffected by, and unaware of, changes occurring at individual source areas<sup>54</sup>. Declining fish stocks, for example, are compensated for by substituting source areas<sup>55</sup>, shifting to new but similar species<sup>54</sup> or replacing wild catch with supply from aquaculture<sup>30</sup>. Similarly, international trade enables countries to displace their land use (for example, deforestation) to other nations<sup>56</sup>. As long as consistent demand exists through globally distributed markets, implementation of policies to mitigate overexploitation in one place-such as protected areas or reduced quotas—may simply increase pressure elsewhere (leakage effects), with a global net decline as a result 10,32.

The current global model of biomass production also spatially decouples consumption from the environmental impacts that it entails  $^{57}$ . These impacts extend beyond direct collateral damages, such as spread of infectious diseases, pollution, habitat degradation and loss of biodiversity. They include reallocation of natural resources (for example, land and water) needed to produce traded commodities destined for direct human consumption or as input to produce biomass with higher protein and nutrient content  $^{6,30,38}$ . The trade of these embodied resources (virtual trade) has been estimated to incorporate 24% and 22% of the global land and water footprint  $^{58}$ , respectively, and account for 11% of global groundwater depletion  $^{59}$ .

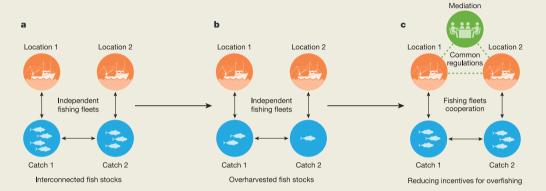
More recently, attention has been given to the way in which decoupling may arise from the growing influence of finance and the emergence of novel financial instruments (Box 1). New types of agricultural insurance have been developed whereby payouts are no longer based on direct measured loss of crops, but are instead triggered by an index, such as a predefined threshold in rainfall  $^{60}$ . Although these index insurance policies present benefits for both insurers (by resolving the problem of moral hazard and reducing the transaction costs of verifying

# Box 2

# A network of networks

The GPE is a worldwide social-ecological network of networks. It is composed of a large number of interacting networks that span sectoral, jurisdictional and geographical boundaries, connect various actors and institutions, and link human societies to the biosphere. Individual nodes in this global network can represent countries, actors, institutions, sectors, species or ecosystems. Links can capture collaboration, trade, policy overlap, environmental effects, species dispersals or trophic interactions<sup>147</sup>. A networkmodelling approach can help to uncover invaluable clues to the resilience of the GPE. For example, the degrees and patterns of connectivity across the many networks can have important shock-amplifying or shock-dampening implications<sup>70,75,77,94,96</sup>.

Another promising, but nascent area of research is the investigation of how well social and ecological systems are aligned within a social-ecological network<sup>148</sup> and empirical assessment of how this social-ecological fit affects outcomes in the system<sup>149</sup>. For example, early results suggest that high levels of social-ecological fit provide an important foundation for more sustainable and adaptive practices to emerge<sup>150</sup> (Box 2 Figure). Future research should aim at exploring the nature and extent of social-ecological fit in the GPE, but also whether enhancing social-ecological fit could lead to mechanisms that unravel and expose different masking effects, such as land displacement, sequential overexploitation and virtual trade.



## Box 2 Figure | Enhanciological fit in fisheries.

a-c, The blue nodes represent interconnected fish stocks (for example, through dispersal and seasonal migration) being targeted at different locations by different fishing fleets (red nodes). For highly mobile fish species such as tuna, the distance between any two localized stocks can be very large and cross several jurisdictional borders. a, If actors (fishing fleets) are operating independently of each other, the ecological and social connectivities are not aligned. This makes it more difficult to respond collectively to changes and disturbances, and can incentivize actors to (knowingly or unknowingly) overharvest fish stocks. b, Overfishing in one location will negatively affect the status of the fish stock in the other location, and vice versa.

c, Bringing the actors closer together (for example, through exchanging information or agreeing on common regulations) would increase the possibilities for tightening social-ecological feedback loops (recoupling), thus reducing incentives for overfishing 149 and increasing the likelihood of the emergence of collective action<sup>147</sup>. Such increased social-ecological fit could be established either by connecting the different actors directly, or indirectly through a third-party mediator. Examples of direct connections include the emergence of fishing collectives<sup>151</sup> or cooperation among global seafood companies<sup>132</sup>. An indirect connection could occur through a regional fishery management organization, in which the participating member states adhere to commonly devised fishery regulations for stocks over which no single state has full authority.

losses) and farmers (by improving access to credit and mitigating climate risk), they are often coupled to the adoption of commercial inputs and specific crops that reinforce the simplification of agricultural landscapes and the homogenization of practices. This increases smallholders' exposure to risks and erodes their ability to adapt to extreme environmental variability  $^{60}$ . Because the actual agricultural performance is no longer relevant for the indemnity payment, farmers are also at risk of experiencing losses but not receiving a payment if the index threshold is not met.

Collectively, these different decoupling mechanisms have ramifications for how production ecosystems and the benefits they produce are perceived, valued and managed 13,61.

## Resilience in the GPE

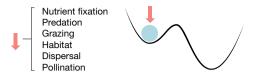
Resilience is a concept that is widely used in science, management and policy. The concept has multiple meanings, which can have consequences for evaluating, understanding, and managing systems,

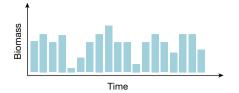
depending on which definition is used. Resilience can refer to the time it takes for a system to return to its original state after perturbation (recovery) or, as in this Perspective, it can describe the extent to which a system can develop with change by absorbing recurrent perturbations, deal with uncertainty and risk, and still sustain its key properties<sup>15</sup>, such as the capacity to feed humanity in the case of the GPE. Concerns have been raised that the profound human influence on the biosphere is eroding resilience and causing abrupt changes in social, ecological and social-ecological systems 62,63. These 'regime shifts' may interact and cascade<sup>64</sup>, thereby producing change at very large scales with severe implications for the wellbeing of human societies<sup>65</sup>. Since the GPE has become a substantial part of the biosphere, investigation of what a hyperconnected, homogenized and decoupled anatomy means for its resilience is urgently needed.

# The structure of fragility

Analysing systems as networks that consist of nodes and links has proved to be a fertile ground for exploring the relationship between

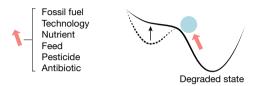
## a Local low-intensity production ecosystem

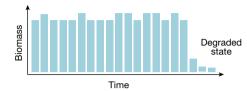




- Artisanal fisheries
- Organic farming
- Free-ranging livestock

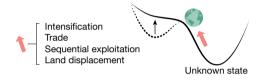
# **b** Local high-intensity production ecosystem

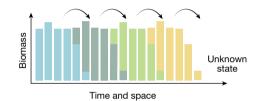




- Industrial fisheries
- Intensive agriculture
- High-density livestock

#### c Global production ecosystem





- Sequential seafood exploitation
- Agricultural boom-and-bust
- Sequential deforestation

**Fig. 2** | **Masking loss of resilience. a**, The state of a local low-intensity production ecosystem (blue dot) is maintained by a suite of biophysical processes (red arrow). Variability in environmental conditions creates fluctuations in biomass output (blue bars). This variability may not be an acceptable solution as drops in the production may not sufficiently meet the needs of people depending on it. **b**, A local high-intensity production ecosystem is kept in a forced state by continuously adding anthropogenic inputs, such as increasing use of antibiotics to avoid diseases in aquaculture and livestock, and herbicides to prevent weeds in crop systems. Intensification increases productivity and suppresses fluctuations in harvestable biomass in the short term (blue bars). This occurs at the expense of eroding resilience in

the long term (dashed line and black arrow), which increases the risk of surpassing a threshold beyond which the system may fall into a degraded state, precipitating a collapse of biomass. **c**, Similarly, the GPE (represented by the Earth) is kept in a forced state through intensification, trade and spatial displacement of activities (red arrow), to maintain a high and predictable global supply of biomass arriving from different stocks, species, geographic locations (multi-coloured bars). Loss of resilience (dashed line and black arrow) is masked at a global level, thus increasing the risk of shifting the GPE into an unknown state. To the right are systems within which examples of the illustrated dynamics can be found (see Supplementary Table 1).

structure and resilience in ecological  $^{66,67}$ , financial  $^{68-70}$ , technological  $^{71,72}$  and climatic systems  $^{73}$ . Depending on how nodes are organized, connectivity can increase or decrease the resilience in a network  $^{74}$ . A recent empirical reconstruction of the global food trade network from 1986 to 2013 showed that it displays characteristics of a heterogeneous network in which countries have many incoming (import) and few outgoing (export) connections, or vice versa  $^{74}$ . The study also found that the food system has become progressively delocalized as a result of globalization (that is, modularity has been reduced). Combining these properties, the authors concluded that resilience in the global food network has declined over the past 20 years and that addition of new trade routes to this heterogeneity will further erode resilience  $^{74}$ .

Another important line of research focuses on the interaction between connectivity and diversity in the network (that is, how nodes are different from each other)<sup>75</sup>. Studies suggest that, in systems in which the diversity of responses among nodes is high and connectivity between them is low, the systemic response to perturbation is gradual. By contrast, if nodes are homogeneous and highly connected, their responses become more synchronized<sup>70,75-77</sup>. The global financial crisis provides an illustrative example in which a small number of tightly connected banks deployed similar risk-management models, thus cultivating homogeneity at the global scale and paving the way for shocks to propagate throughout the financial system<sup>68</sup>. Connectivity and diversity therefore determine whether a system has a shock-dampening or a shock-amplifying effect when exposed to perturbations<sup>77</sup>. Linking

networks together can help to reduce pressure in individual networks, but may occur at the expense of increasing fragility of the broader interconnected network  $^{71}$ . Indeed, studies in power-communication  $^{72}$ , financial  $^{68}$  and ecological  $^{64}$  systems have shown that a large interconnected 'network of networks' can be intrinsically more fragile than each network in isolation.

In the GPE, intensification and globalization have produced strong interdependencies within and among sectors. In parallel, homogenization has reduced the diversity of ways in which species, people, sectors and institutions can respond to change (loss of response diversity) $^{78}$  as well as their potential to functionally complement each other (loss of redundancy) $^{16,79}$ . This suggests that the GPE possesses features that could amplify shocks $^{80}$ . Understanding such potential shock-amplifying behaviour will require a better evaluation of how ecological, social and social-ecological connectivity and diversity interact (Box 2).

# Masking loss of resilience

Fluctuations in harvestable biomass outputs influence producer income and undermine the continued and stable supply to consumers (Fig. 2a). Strategies that reduce this variation to improve efficiency and predictability are therefore frequently sought. However, enhanced short-term control can have implications for resilience in the long term.

Increasing variability (variance) can be a signal of declining resilience in complex systems, including ecosystems and social-ecological systems<sup>75</sup> (but see ref. <sup>81</sup>). Hence, intensification strategies that deliberately

# Box 3

# Linking anatomy to resilience: empirical examples from food production

The resilience of a given system depends on the interplay between exogenous (external) forces and the endogenous (internal) properties of the system (Box 3 Figure). There is growing evidence that key external and internal dynamics of the GPE have been altered to the extent that they are compromising its resilience. Here we consider and empirically illustrate two examples of such dynamics.

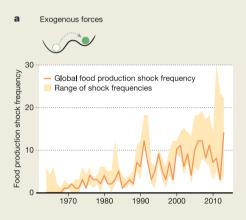
## Altered perturbations and increasing food-shock frequency

As connectivity and homogeneity increase, shocks within a geographic area or sector can become globally contagious and more prevalent<sup>97</sup>. Major drivers of these shocks include extreme weather events, spread of disease and geopolitical or economic conflicts, although their relative importance can differ across regions. For instance, droughts and floods have been particularly dominant forces of sudden declines in crop production over the past decades in South Asia, whereas geopolitical and economic crises are leading drivers of agricultural shocks in sub-Saharan Africa80. Furthermore, geopolitical and economic events—such as withdrawal of subsidies, reduced export markets and internal conflicts—tend to generate shocks that span multiple sectors across both land and sea80. Importantly, despite efforts to maintain high yields, food production shocks have become more frequent over the past 50 years (Box 3 Fig. a). These shocks pose a threat to food

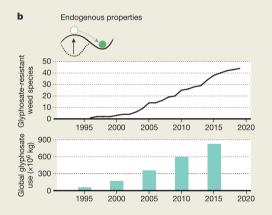
security and the resilience of the global food system through price volatilities and effects on trade<sup>80</sup>.

#### The looming threat of herbicide resistance

Homogenization towards pesticide-intensive production practices in agroecosystems has increased selection for pesticide resistance and reduced the resilience of the GPE to pests and pathogens<sup>137</sup>. For instance, according to The International Survey of Herbicide Resistant Weeds (www.weedscience.org), as of 2019, weeds have evolved resistance to 167 different herbicides and to 23 of the 26 known herbicide sites of action. Glyphosate is currently the most commonly applied weedkiller, accounting for approximately twothirds of all herbicide use globally. First introduced in 1974, the application of glyphosate began to escalate in the mid-1990s with the rapid adoption of glyphosate-resistant transgenic crops, which enabled farmers to use glyphosate liberally as a strategy to maintain and enhance yields (Box 3 Fig. b). However, such short-term damping of variability can erode resilience in the long term<sup>82</sup>. The overreliance on glyphosate rapidly accelerated the evolution and spread of glyphosate-resistant weeds across all major economies (Box 3 Fig. b), ultimately creating conditions for a looming global failure of weed management 138.



Box 3 Figure | Linking anatomy to resilience. a, b, The schematic of the stability landscape illustrates two ways by which exogenous forces (a) and changes in endogenous system properties (b) can lead to state shifts (dashed arrows) in food production systems. a, The annual increase in global food production shock frequency, including fisheries, aquaculture, crop and livestock sectors.



The confidence interval (orange shaded region) represents the range of plausible shock frequencies under different model parameters used for shock detection80. b, The growing number of glyphosate-resistant weeds (black line) as global glyphosate use (blue bars) increases. See Supplementary Note 1 for methodology and data sources.

suppress variance may remove a useful warning of declining resilience in production ecosystems, sectors and the broader GPE<sup>82</sup>. Variance is often suppressed by controlling stress and stochastic perturbations such as grazing, fire and pest outbreaks. Such events have been proposed to increase system resilience in the long term by selecting for particular tolerant genes, species traits or practices 83,84. Therefore, preventing these events may gradually erode resilience, making systems more vulnerable to disturbances that could previously be absorbed. Suppressing short-term variance can also lead to an accumulation of variance in the longer term<sup>82</sup>. As variance accumulates, more force (that is, human input) is required to maintain the system in a desired state (Fig. 2b). Resilience under such conditions has been described as 'coerced'<sup>2</sup>. In forest production ecosystems, for example, stochastic wildfires are often curbed to maintain high and stable yields of harvestable biomass. However, small-scale fires have an important role in reducing the accumulation of dead wood and creating a heterogeneity of patches with less-flammable species that reduces the risk of ignition and prevents fires from propagating 85,86. This allows the system to suffer fire without eminent risk of crossing a critical threshold whereby it becomes catastrophic and uncontrollable. By contrast, when

small-scale wildfires are suppressed, homogeneity increases and the amount of wood fuel piles up. This creates a situation in which a single ignition could potentially set the whole forest on fire. If a catastrophic fire unfolds, it can start to interact with the atmosphere and generate convection-driven winds, which further increase its size, spread and speed, making the fire unstoppable St. Consequently, management aimed at controlling short-term variability breeds systemic vulnerability in the long run  $(80 \times 3)$ .

The anatomy of the GPE provides for spatial suppression of and accumulation of variance at a global level because components of the system (for example, sectors, places and stocks) are often viewed and governed in isolation (Fig. 2c). This global coercion of resilience is facilitated by sequential exploitation and displacement of activities. For example, countries transitioning from net deforestation to net reforestation may do so through geographic substitution  $^{87}$ . In Vietnam, forest cover increase was achieved at the expense of deforestation in neighbouring countries such as Cambodia and Laos  $^{56}$ . Similarly, following the collapse of the North Sea cod ( $Gadus\,morhua$ ) population, UK imports shifted to Atlantic cod sourced from Iceland and the Faeroes, leaving UK consumers relatively unaffected and unaware  $^{54}$ . Such decoupling mechanisms could explain why national supply stability tends to increase as countries' reliance on trade grows  $^{88}$ , although it may contribute to global instability in the long term  $^{74}$ .

# Altered disturbance landscape and systemic risks

Resilience management has generally focused on local systems and their capacity to deal with a narrow range of well-known shocks<sup>89</sup>, such as drought, fire, pest outbreaks and, increasingly, climate change. However, perturbations that previously had only minor or no effects on a certain production ecosystem may suddenly become significant as sectors are progressively intensified and intertwined. For example, droughts or crop pest outbreaks may cause disruption in seafood production, as the aquaculture sector is now dependent on agriculture for crop feeds<sup>30</sup>. Moreover, the GPE has become increasingly exposed to price fluctuation in inputs (for example, fossil fuels, fertilizers and technology)<sup>90</sup>, shifts in global consumer preferences (for example, diets)91, changes in policies (for example, regulations on energy and exports)92 and speculation on food commodities93. Concerns have also been raised about the vulnerability of the infrastructure network on which trade of biomass relies, such as choke points in maritime transportation, which could generate significant instabilities if disrupted 90,94.

As connectivity and homogeneity increase, shocks that were previously contained within a geographic area or a sector are becoming globally contagious and more prevalent (Box 3). For example, protectionist trade strategies, such as implementation of export bans following droughts to protect populations in producing countries, can affect nations that rely on trade to balance their food needs<sup>52,95</sup>. Interest in these types of interconnected risks has increased in recent years along with the terminology to describe them, including nested and teleconnected vulnerabilities<sup>12</sup>, hyper-risks<sup>76</sup>, femtorisks<sup>96</sup>, global systemic risks<sup>94</sup> and Anthropocene risk<sup>97</sup>. They stem from interactions at the interface of multiple systems (for example, climatic, ecological, political, financial and technological), making causal links opaque and outcomes difficult to foresee<sup>76</sup>.

Despite the inherent uncertainties, this broad spectrum of perturbations and interconnected shocks must be considered to adequately manage resilience in the GPE. It also suggests that the limits of the GPE in satisfying demands for harvestable biomass may be set by the potential consequences of these emergent risks, as opposed to hard upper limits to production per se. The future will require confronting risks that we know little about <sup>89,98</sup>, such as the consequences of an expanding global financial sector (Box 1) and new technologies, including the growth of genetic engineering and synthetic biology <sup>42</sup>. It will also entail accounting for interactions with non-biomass producing sectors that, for instance, support critical infrastructure or energy supply.

Competition between production sectors for land and resources (most importantly water) is indeed likely to intensify as demand continues to grow and effects of climate change unfold.

## Towards a sustainable GPE

Providing a growing human population with food, fibre and fuel in a sustainable and fair way is one of the grand challenges facing humanity. Although the GPE has offered huge benefits by increasing the production of certain desired species 99, the intensification and simplification of production ecosystems have been criticized from ecological 2.6.11, social 17.39 and social-ecological perspectives 100. Consequently, we argue that it should be substantially and deliberately transformed towards a sustainable trajectory, on which: (1) the demands for biomass are met in a fair and just way, without undermining the functioning of the biosphere, (2) connectivity is capitalized on to improve sustainability, (3) biological and social diversity is enhanced to ensure building blocks for adaptability and transformation in the face of change, and (4) feedback loops are strengthened (recoupled) to avoid masking effects and coercion of resilience.

Determining the boundary conditions characterizing a sustainable GPE is a challenging task that will involve a mix of approaches. The planetary boundaries framework <sup>101</sup> can be used to define global and regional limits in biophysical processes—the 'safe operating environmental space'—that must not be transgressed if humanity is to stay away from systemic and potentially irreversible shifts in the biosphere. For example, this framework was recently applied to quantitatively estimate how to keep the global food system within environmental limits<sup>102</sup>. Combined with the aspirational social goals framework ('safe and just space for humanity')<sup>103</sup>, this can provide a starting point for discussions around levels of acceptable risk and trade-offs between productivity, sustainability and equity<sup>104</sup>.

Steering the GPE towards a sustainable trajectory will also require a combination of more specific strategies and solutions, as well as careful consideration of their feasibility and the trade-offs involved. Although the polarized debate between the integration (land sharing) and separation (land sparing) of conservation and production fits into discussions around food production and land scarcity, it is ill-suited to address issues of scale (for example, temporal variation in agricultural land use patterns and total area for conservation) or effects of globalization (for example, displacement activities)<sup>105</sup>. The land-sparing versus landsharing debate is too often framed as a binary choice, ignoring possible middle ground and cross-fertilization. Within this context, sustainable intensification has gained momentum in discussions around global sustainability and has become a policy goal for many institutions to deliver on global social and environmental commitments (for example, the UN Sustainable Development Goals and the Paris Agreement). However, it has also been criticized for having a narrow focus on efficiency gains and technological interventions 106. More systemic forms of sustainable intensification have therefore begun to occur at large scales and across a wide range of agroecosystems, to redesign the composition and structure of production ecosystems and harness a broad range of ecological processes such as predation, parasitism, herbivory, nitrogen fixation and pollination<sup>107</sup>. Further efforts towards a sustainable GPE include approaches to ensure more stable food supplies by increasing national crop diversity<sup>108</sup>, broad-scale shifts in diets and strategies to  $reduce \,loss\, and\, waste\, of\, biomass^{102}, and\, the\, integration\, of\, local\, realities$ and contexts, such as procedural justice and equitable distribution of benefits from multi-functional land and seascapes 17,109.

Although these initiatives are contributing to sustainability in important ways, they are being challenged by an expanding GPE in which systemic, sectoral and jurisdictional boundaries are increasingly blurred. Acting on this new reality entails creating conditions that foster innovation, incentivize transformation and encourage new partnerships across different sectors and actor groups<sup>110</sup>. For this reason, we

propose three entry points towards a more sustainable GPE that have great transformative potential but are still in their infancy.

# Redirecting finance for sustainability

Financial investments—public or private—are increasingly recognized as key leverage points for achieving sustainability<sup>111-113</sup>. Government subsidies channel large amounts of public capital into the different sectors of the GPE, ultimately influencing practices and species production on the ground. Whereas subsidies have mostly been associated with unsustainable practices, such as fuelling over-capacity in the fishing industry<sup>114</sup>, they could also provide powerful incentives for improved sustainability if linked to the right criteria. In the European Union's reformed Common Agricultural Policy, for example, a direct payment scheme is used to incentivize sustainable resource management, in which farmers who comply with greening measures (that is, addressing biodiversity loss, avoiding crop monoculture and securing carbon sequestration) benefit financially from payments (but see ref. 115). Another recent government-led action is the alliance of Central Banks and Supervisors Network for Greening the Financial System (NGFS), formed during the One Planet Summit in 2017 to explore the role of and possibilities for central banks to use their mandate to incentivize economies to transition to more sustainable pathways116.

Private financial actors such as asset managers and commercial banks channel the bulk of capital behind the expansion of the GPE by investing in or lending to companies in different production sectors. Although direct causality between financial flows and environmental change is often opaque<sup>112,117</sup>, such investments represent a potential source of influence over corporate practices. Shareholders of publicly listed companies have the ability to affect a firm's sustainability performance by exercising their voting rights at shareholder meetings (shareholder activism). They can engage directly with the corporate leadership on governance and policy, or indirectly through chains of ownership and threats of divestment. For example, the world's largest sovereign wealth fund-Norway's Government Pension Fund-has divested from 32 companies involved in unsustainable palm oil production since deforestation became an ethical criterion in 2012 (https://www.nbim.no). The insurance sector could also provide important leverage towards more sustainable practices-for instance, by refusing to insure fishing vessels associated with illegal, unreported and unregulated fishing<sup>118</sup>. Similarly, loan covenants (that is, the specific conditions associated with credit lending) provide a powerful tool for banks to influence the behaviour of borrowing companies operating in the GPE by denying access to clients that do not comply with sustainability standards and providing incentives so that better sustainability performance results in reduced interest rates 113. In this context, pressure from governments and finance ministries will be essential to promote new norms and regulations that can align banks, financial markets and other investors with sustainability goals<sup>113</sup>.

# Radical transparency and traceability

Consumers can be influential in promoting sustainability by aligning their purchasing with sustainable thinking. They are also important as citizens whose perceptions and opinions drive the political will to address sustainability issues. Education and provision of informationsuch as certification, labelling schemes and public campaigns—are therefore central instruments for consumers to make informed decisions<sup>54</sup>. However, if as a society we do not know where, how, in what quantity and by whom a given commodity is produced, it is arguably difficult to tackle sustainability challenges119.

Whereas transparency is necessary to assess the environmental sustainability of corporate and financial activities, traceability represents a key mechanism by which corporations can ensure that their supply chains are devoid of unacceptable behaviour, ranging from illegal sourcing and forced labour to poor sanitation and mislabel $ling^{120-122}. \, Many \, of the \, operations \, of the \, corporate \, and \, financial \, world \,$  are still plagued by opaqueness<sup>119,123</sup>, including secrecy around financial transactions and corporate loans<sup>117</sup>, as well as poor disclosure on implementation of corporate policy and internal allocation of capital (see https://www.ifrs.org/).

Radical transparency and traceability require the disclosure of production volumes and practices. It also demands that corporate and governmental policies are put in place to ensure that social and environmental criteria are met in all supply chain segments, as well as mechanisms to monitor how such regulations are implemented and enforced<sup>124,125</sup>. To date, improved corporate disclosure has largely been driven by voluntary action 125 under the scrutiny of non-governmental organizations (NGOs). Whereas the Global Reporting Initiative (https:// www.globalreporting.org) is a prominent example of widely adopted sustainability reporting standards, the more recent World Benchmarking Alliance (https://www.worldbenchmarkingalliance.org/) encourages companies to disclose information that allows evaluation of their operations in relation to industry benchmarks. Even though mandatory reporting is increasing globally, limited regulation contributes to poor transparency and sustainability-related corporate reporting remains voluntary in many jurisdictions (see https://www.carrotsandsticks. net). More stringent and clearly articulated criteria for disclosure therefore represent an important step towards more transparent corporate practices.

Emerging digital technologies that deliver decentralized systems, such as blockchain 126,127, could resolve some of these issues and improve traceability in the GPE. However, these technologies are energy-intensive and interoperability remains a hurdle because seamless communication between digital platforms and agreed-on data for transmission are largely unrealized 128. Thus, barriers to chain-wide traceability are not just technological but also organizational, and will require changes in legislation and the institutions that govern trade to stimulate cooperation throughout supply chains<sup>124</sup>.

# Keystone actors as global agents of change

A key facet of sustainability science is that the identification of challenges and their solutions requires collaboration between researchers and actors from outside academia<sup>129</sup>. Generally, these actors encompass local communities, indigenous groups, management agencies, NGOs and government actors. More recently, however, increasing attention has been directed towards large transnational corporations and their role as a threat to, or as an opportunity for, sustainable transformation  $^{37,130,131}$ .

Private governance raises concerns associated with accountability. fair representation and global equity<sup>36</sup>. Nevertheless, transnational corporations have become a central feature of the GPE (that is, keystone actors), with a capacity to influence practices across supply chains and geographical locations<sup>35</sup>, and thus have the potential to become powerful agents of change for improved sustainability<sup>37</sup>. An increasing number of private sector initiatives is emerging with the intention to mobilize companies to take tangible actions, make investments and form partnerships to deliver on sustainability<sup>37</sup>.

Scientists have an important role to play in this context, acting as independent knowledge brokers to ensure that the agendas of keystone actors are based on scientific evidence and align with long-term sustainability goals. Seafood Business for Ocean Stewardship (SeaBOS) provides an unconventional example of a co-production initiative in which scientists directly engaged with the world's largest seafood companies to stimulate transformative change towards improved ocean stewardship<sup>132</sup>. Drawing on an empirical identification of the largest companies involved in aquaculture and wild-capture fisheries<sup>35</sup> this global science-business initiative emerged in 2016 with a number of task forces led by member companies in collaboration with and supported by scientists (https://www.keystonedialogues.earth). While the long-term outcome remains to be evaluated, the 10 companies engaged in SeaBOS can influence the strategic direction of more than 600 subsidiaries with operations in at least 90 different countries<sup>132</sup>.

Although this presents a promising approach to be replicated in other sectors in the GPE, such engagements do not come without risk. For scientists, they may cause reputational damage and loss of credibility if companies use the initiative for greenwashing purposes or if they fall short on their promises. For the private sector, they may lead to competitive disadvantage and loss of profit in the short term if other companies do not participate. Nevertheless, with renewable biomass and global sustainability at stake, there are strong incentives for novel science—business partnerships to emerge in combination with effective public policies and improved governmental regulations<sup>37</sup>.

## Conclusion

The rate of change of the Earth's system is accelerating. Unless meaningful actions are taken within the next decade, we will almost certainly face a changed and increasingly unstable climate regime<sup>65</sup>, with serious disruptions to the GPE as a consequence. The current GPE is itself a major driver of this change, accounting for nearly a quarter of all anthropogenic greenhouse gas emissions over the past decade<sup>133</sup>. As a result, agriculture, forestry and fishing are increasingly embedded in international efforts to tackle climate change. Government policies are essential to foster such transformations and align the global economy with sustainability goals. In the face of the urgency and complexity of this challenge, we also need to explore new spaces for innovation and transformation. Although the avenues proposed here are in their infancy, they provide potential entry points for transformative change and a complement to effective governmental regulations. Ultimately, moving towards a more sustainable GPE is likely to require radical shifts in deeply held values, education systems and social behaviour that underpin current economic paradigms, consumption patterns and power relationships<sup>134–136</sup>. Scientists have an important role to play in this process.

- Bennett, E. M. & Balvanera, P. The future of SMEs in a globalized world. Front. Ecol. Environ. 5, 191–198 (2007)
- Rist, L. et al. Applying resilience thinking to production ecosystems. Ecosphere 5, 1–11 (2014).
  - This study shows how anthropogenic inputs of external resources can lead to a 'coercion' of resilience and how the global connectivity among production ecosystems can obscure signals indicating resilience loss.
- Vitousek, P. M., Mooney, H. A., Lubchenco, J. & Melillo, J. M. Human domination of Earth's ecosystems. Science 277, 494–499 (1997).
- Barnosky, A. D. et al. Approaching a state shift in Earth's biosphere. Nature 486, 52–58 (2012).
- Ellis, E. C. Anthropogenic transformation of the terrestrial biosphere. Philos. Trans. R. Soc. A 369, 1010–1035 (2011).
  - The paper shows how humans have transformed the biosphere into intensified anthropogenic biomes over the past century.
- 6. Foley, J. A. et al. Solutions for a cultivated planet. Nature 478, 337–342 (2011).
- Gauthier, S., Bernier, P., Kuuluvainen, T., Shvidenko, A. Z. & Schepaschenko, D. G. Boreal forest health and global change. Science 349, 819–822 (2015).
- The State of World Fisheries and Aquaculture 2018—Meeting the Sustainable Development (FAO, 2018).
- Lester, S. E. et al. Marine spatial planning makes room for offshore aquaculture in crowded coastal waters. Nat. Commun. 9, 945 (2018).
- Lambin, E. F. & Meyfroidt, P. Global land use change, economic globalization, and the looming land scarcity. Proc. Natl Acad. Sci. USA 108, 3465–3472 (2011).
   Tilman, D. Cassman, K. G. Matson, P. A. Naylor, R. & Polasky, S. Adricultural sustainabil
- Tilman, D., Cassman, K. G., Matson, P. A., Naylor, R. & Polasky, S. Agricultural sustainability and intensive production practices. *Nature* 418, 671–677 (2002).
- Adger, W. N. Eakin, H. & Winkels, A. Nested and teleconnected vulnerabilities to environmental change. Front. Ecol. Environ. 7, 150–157 (2009).
- 13. Folke, C. et al. Reconnecting to the biosphere. Ambio 40, 719-738 (2011).
- Liu, J. et al. Systems integration for global sustainability. Science 347, 1258832 (2015).
   This paper suggests ways forward for improved integration of distal human and natural components to address global sustainability.
- Folke, C. et al. Social-ecological resilience and biosphere-based sustainability science. Ecol. Soc. 21, art41 (2016).
  - This work shows how resilience can be used as a lens to understand social-ecological systems and to address biosphere-based sustainability.
- Biggs, R. et al. Toward principles for enhancing the resilience of ecosystem services. Annu. Rev. Environ. Resour. 37, 421–448 (2012).
- Fischer, J., Meacham, M. & Queiroz, C. A plea for multifunctional landscapes. Front. Ecol. Environ. 15, 59 (2017).
- Steffen, W., Broadgate, W., Deutsch, L., Gaffney, O. & Ludwig, C. The trajectory of the Anthropocene: the Great Acceleration. Anthr. Rev. 2, 81–98 (2015).

- de Vrese, P., Hagemann, S. & Claussen, M. Asian irrigation, African rain: remote impacts of irrigation. Geophys. Res. Lett. 43, 3737–3745 (2016).
- Bonan, G. B. Forests and climate change: forcings, feedbacks, and the climate benefits of forests. Science 320, 1444–1449 (2008).
- Barange, M. et al. Impacts of Climate Change on Fisheries and Aquaculture: Synthesis of Current Knowledge, Adaptation and Mitigation Options (FAO, 2018).
- Anderson, K. Globalization's effects on world agricultural trade, 1960–2050. Philos. Trans. R. Soc. Lond. B 365, 3007–3021 (2010).
- Marchand, P. et al. Reserves and trade jointly determine exposure to food supply shocks. Environ. Res. Lett. 11. 095009 (2016).
- D'Odorico, P., Carr, J. A., Laio, F., Ridolfi, L. & Vandoni, S. Feeding humanity through global food trade. Earth's Future 2, 458–469 (2014).
- Weinzettel, J., Hertwich, E. G., Peters, G. P., Steen-Olsen, K. & Galli, A. Affluence drives the global displacement of land use. Glob. Environ. Change 23, 433–438 (2013)
- 26. D'Odorico, P. & Rulli, M. C. The fourth food revolution. Nat. Geosci. 6, 417–418 (2013).
- Regional Trade Agreements Information System Database (WTO, accessed 14 May 2018); https://rtais.wto.org/UI/PublicMaintainRTAHome.aspx.
- Kastner, T., Erb, K. H. & Haberl, H. Rapid growth in agricultural trade: Effects on global area efficiency and the role of management. Environ. Res. Lett. 9, 034015 (2014).
- Rudel, T. K. et al. Agricultural intensification and changes in cultivated areas, 1970–2005.
   Proc. Natl Acad. Sci. USA 106, 20675–20680 (2009).
- Troell, M. et al. Does aquaculture add resilience to the global food system? Proc. Natl Acad. Sci. USA 111, 13257-13263 (2014).
- MacDonald, G. K. et al. Rethinking agricultural trade relationships in an era of globalization. Bioscience 65, 275–289 (2015).
- Fuchs, R. et al. Why the US-China trade war spells disaster for the Amazon. Nature 567, 451-454 (2019).
- Galloway, J. N. et al. International trade in meat: the tip of the pork chop. Ambio 36, 622–629 (2007).
- Fry, J. P. et al. Environmental health impacts of feeding crops to farmed fish. Environ. Int. 91, 201–214 (2016).
- Österblom, H. et al. Transnational corporations as 'keystone actors' in marine ecosystems. PLoS ONE 10, e0127533 (2015).
- Clapp, J. & Fuchs, D. A. (eds) Corporate Power in Global Agrifood Governance (MIT Press, 2009)
- Folke, C. et al. Transnational corporations and the challenge of biosphere stewardship. Nat. Ecol. Evol. 3, 1396–1403 (2019).
- Naylor, R. et al. Losing the links between livestock and land. Science 310, 1621–1622 (2005).
- Hendrickson, M. K. Resilience in a concentrated and consolidated food system. J. Environ. Stud. Sci. 5, 418–431 (2015).
- Jefferson, O. A., Köllhofer, D., Ehrich, T. H. & Jefferson, R. A. The ownership question of plant gene and genome intellectual properties. Nat. Biotechnol. 33, 1138–1143 (2015).
- Baiser, B., Olden, J. D., Record, S., Lockwood, J. L. & McKinney, M. L. Pattern and process of biotic homogenization in the New Pangaea. Proc. R. Soc. Lond. B 279, 4772–4777 (2012).
- Carroll, S. P. et al. Applying evolutionary biology to address global challenges. Science 346, 1245993 (2014).
- Koh, L. P. & Wilcove, D. S. Is oil palm agriculture really destroying tropical biodiversity? Conserv. Lett. 1, 60–64 (2008).
- Gómez-González, S., Ojeda, F. & Fernandes, P. M. Portugal and Chile: longing for sustainable forestry while rising from the ashes. Environ. Sci. Policy 81, 104–107 (2018).
- Gossner, M. M. et al. Land-use intensification causes multitrophic homogenization of grassland communities. Nature 540, 266–269 (2016).
- Rodrigues, J. L. M. et al. Conversion of the Amazon rainforest to agriculture results in biotic homogenization of soil bacterial communities. *Proc. Natl Acad. Sci. USA* 110, 988–993 (2013).
- Magurran, A. E., Dornelas, M., Moyes, F., Gotelli, N. J. & McGill, B. Rapid biotic homogenization of marine fish assemblages. Nat. Commun. 6, 8405 (2015).
- 48. FAOSTAT Statistics Database (FAO, 2017).
- Khoury, C. K. et al. Increasing homogeneity in global food supplies and the implications for food security. Proc. Natl Acad. Sci. USA 111, 4001–4006 (2014).
  - This paper shows that national portfolios of food supplies have seen increased crop species diversity, whereas globally, they have become more homogeneous in composition.
- Borras, S. M., Franco, J. C., Isakson, S. R., Levidow, L. & Vervest, P. The rise of flex crops and commodities: implications for research. J. Peasant Stud. 43, 93–115 (2016).
- 51. Foley, J. A. et al. Global consequences of land use. Science 309, 570-574 (2005).
- Fader, M., Gerten, D., Krause, M., Lucht, W. & Cramer, W. Spatial decoupling of agricultural production and consumption: quantifying dependences of countries on food imports due to domestic land and water constraints. *Environ. Res. Lett.* 8, 021002 (2013).
- Clapp, J. Financialization, distance and global food politics. J. Peasant Stud. 41, 797–814 (2014).
  - A study that provides a new perspective on the increasing role of finance in an intensified global food system and how it weakens feedback effects between production and consumption.
- Crona, B. I. et al. Masked, diluted and drowned out: how global seafood trade weakens signals from marine ecosystems. Fish. 17, 1175–1182 (2016).
- Berkes, F. et al. Globalization, roving bandits, and marine resources. Science 311, 1557–1558 (2006).
- Meyfroidt, P. & Lambin, E. F. Forest transition in Vietnam and displacement of deforestation abroad. Proc. Natl Acad. Sci. USA 106, 16139–16144 (2009).
- Srinivasan, U. T. et al. The debt of nations and the distribution of ecological impacts from human activities. Proc. Natl Acad. Sci. USA 105, 1768–1773 (2008).
- Hoekstra, A. Y. & Wiedmann, T. O. Humanity's unsustainable environmental footprint. Science 344, 1114–1117 (2014).

- Dalin, C., Wada, Y., Kastner, T. & Puma, M. J. Groundwater depletion embedded in international food trade. Nature 543, 700-704 (2017).
- 60. Isakson, S. R. Derivatives for development? Small-farmer vulnerability and the financialization of climate risk management. J. Agrar. Change 15, 569-580 (2015).
- Steneck, R. S. et al. Creation of a gilded trap by the high economic value of the Maine lobster fishery. Conserv. Biol. 25, 904-912 (2011).
- Hughes, T. P., Carpenter, S., Rockström, J., Scheffer, M. & Walker, B. Multiscale regime shifts and planetary boundaries. Trends Ecol. Evol. 28, 389-395 (2013).
- Rocha, J. C., Peterson, G. D. & Biggs, R. Regime shifts in the anthropocene: drivers, risks, 63. and resilience. PLoS ONE 10, e0134639 (2015).
- Rocha, J. C., Peterson, G., Bodin, Ö. & Levin, S. Cascading regime shifts within and across scales. Science 362, 1379-1383 (2018).
- 65. Steffen, W. et al. Trajectories of the Earth System in the Anthropocene. Proc. Natl Acad. Sci. USA 115, 8252-8259 (2018).
- van Nes, E. H. & Scheffer, M. Implications of spatial heterogeneity for catastrophic regime 66. shifts in ecosystems. Ecology 86, 1797-1807 (2005).
- 67. Thébault, E. & Fontaine, C. Stability of ecological communities and the architecture of mutualistic and trophic networks. Science 329, 853-856 (2010).
- 68. Haldane, A. G. & May, R. M. Systemic risk in banking ecosystems. Nature 469, 351-355 (2011).
- 69 Bardoscia, M., Battiston, S., Caccioli, F. & Caldarelli, G. Pathways towards instability in financial networks, Nat. Commun. 8, 14416 (2017)
- Battiston, S. et al. Complexity theory and financial regulation. Science 351, 818-819 70 (2016)
- 71. Brummitt, C. D., D'Souza, R. M. & Leicht, E. A. Suppressing cascades of load in interdependent networks. Proc. Natl Acad. Sci. USA 109, E680-E689 (2012).
- Buldyrev, S. V., Parshani, R., Paul, G., Stanley, H. E. & Havlin, S. Catastrophic cascade of 72 failures in interdependent networks. Nature 464, 1025-1028 (2010).
- 73. Donges, J. F., Schultz, H. C. H., Marwan, N., Zou, Y. & Kurths, J. Investigating the topology of interacting networks. Eur. Phys. J. B 84, 635-651 (2011).
- Tu, C., Suweis, S. & D'Odorico, P. Impact of globalization on the resilience and sustainability of natural resources. Nat. Sustain. 2, 283-289 (2019). This study explores how resilience in a system may either increase or decrease depending on the network structure, and shows that the resilience of the global food trade network has decreased over the past 20 years.
- Scheffer, M. et al. Anticipating critical transitions. Science 338, 344-348 (2012).
- 76. Helbing, D. Globally networked risks and how to respond. Nature 497, 51-59 (2013). This work shows how an increasingly complex and networked world paves the way for
  - risks to emerge and interact, while at the same time reducing our ability to understand and manage these risks.
- 77. Haldane, A. G. Rethinking the financial network https://www.bankofengland.co.uk/ speech/2009/rethinking-the-financial-network (Bank of England, 2009).
- 78. Elmqvist, T. et al. Response diversity, ecosystem change, and resilience. Front. Ecol. Environ 1 488-494 (2003)
- Grêt-Regamey, A., Huber, S. H. & Huber, R. Actors' diversity and the resilience of socialecological systems to global change. Nat. Sustain. 2, 290-297 (2019).
- 80. Cottrell, R. S. et al. Food production shocks across land and sea. Nat. Sustain. 2, 130-137 (2019)
  - This study shows how the frequency of food production shocks has increased in all major food sectors across land and sea over the past 53 years.
- Boettiger, C., Ross, N. & Hastings, A. Early warning signals: The charted and uncharted territories. Theor. Ecol. 6, 255-264 (2013).
- Carpenter, S. R., Brock, W. A., Folke, C., van Nes, E. H. & Scheffer, M. Allowing variance may enlarge the safe operating space for exploited ecosystems. Proc. Natl Acad. Sci. USA 112, 14384-14389 (2015).
  - This paper shows how management of short-term variance increases the risk of crossing critical ecosystem thresholds in the long term, resulting in less desirable ecosystem states
- Seidl, R., Rammer, W. & Spies, T. A. Disturbance legacies increase the resilience of forest ecosystem structure, composition, and functioning. Ecol. Appl. 24, 2063-2077 (2014).
- Leslie, P. & McCabe, J. T. Response diversity and resilience in social-ecological systems. Curr. Anthropol. 54, 114-143 (2013).
- 85 Peters, D. P. C. et al. Cross-scale interactions, nonlinearities, and forecasting catastrophic events. Proc. Natl Acad. Sci. USA 101, 15130-15135 (2004).
- 86. Stephens, S. L. et al. Temperate and boreal forest mega-fires: Characteristics and challenges, Front, Ecol, Environ, 12, 115-122 (2014).
- 87. Meyfroidt, P., Rudel, T. K. & Lambin, E. F. Forest transitions, trade, and the global displacement of land use, Proc. Natl Acad. Sci. USA 107, 20917-20922 (2010).
- Sartori, M. & Schiavo, S. Connected we stand: A network perspective on trade and global 88. food security. Food Policy 57, 114-127 (2015).
- 89. Folke, C. et al. Resilience thinking: Integrating resilience, adaptability and transformability. Ecol. Soc. 15, 20-28 (2010).
- 90 Homer-Dixon, T. et al. Synchronous failure: The emerging causal architecture of global crisis, Ecol. Soc. 20, 6 (2015).
- Kearney, J. Food consumption trends and drivers. Philos. Trans. R. Soc. B 365, 2793–2807 91. (2010).
- Banse, M., Van Meijl, H., Tabeau, A. & Woltjer, G. Will EU biofuel policies affect global 92. agricultural markets? Eur. Rev. Agric. Econ. 35, 117-141 (2008).
- 93. Colbran, N. The financialisation of agricultural commodity futures trading and its impact on the 2006-2008 global food crisis. In 3rd Bienn. Ingram Colloq. Int. Law Dev. Univ. New South Wales 1-13 (2011).
- Centeno, M. A., Nag, M., Patterson, T. S., Shaver, A. & Windawi, A. J. The emergence of global systemic risk. Annu. Rev. Sociol. 41, 65-85 (2015).
- Wood, S. A., Smith, M. R., Fanzo, J., Remans, R. & DeFries, R. S. Trade and the equitability of global food nutrient distribution. Nat. Sustain. 1, 34-37 (2018).
- Frank, A. B. et al. Dealing with femtorisks in international relations. Proc. Natl Acad. Sci. USA 111, 17356-17362 (2014).

- Keys, P. W. et al. Anthropocene risk, Nat. Sustain, 2, 667-673 (2019).
- Carpenter, S. R. et al. General resilience to cope with extreme events. Sustainability 4, 98. 3248-3259 (2012)
- Kareiva, P., Watts, S., McDonald, R. & Boucher, T. Domesticated nature: shaping 99. landscapes and ecosystems for human welfare. Science 316, 1866-1869 (2007).
- 100. Rasmussen, L. V. et al. Social-ecological outcomes of agricultural intensification. Nat. Sustain. 1, 275 (2018).
- 101. Steffen, W. et al. Planetary boundaries: guiding human development on a changing planet. Science 347, 1259855 (2015).
- 102. Springmann, M. et al. Options for keeping the food system within environmental limits. Nature 562, 519-525 (2018).
- 103. Raworth, K. Doughnut economics: seven ways to think like a 21st-century economist. (Chelsea Green Publishing, 2017).
- 104. O'Neill, D. W., Fanning, A. L., Lamb, W. F. & Steinberger, J. K. A good life for all within planetary boundaries. Nat. Sustain. 1, 88-95 (2018).
- 105. Fischer, J. et al. Land sparing versus land sharing; moving forward, Conserv. Lett. 7. 149-157 (2014).
- 106. Loos, J. et al. Putting meaning back into 'sustainable intensification'. Front. Ecol. Environ. 12. 356-361 (2014).
- 107. Pretty, J. et al. Global assessment of agricultural system redesign for sustainable intensification, Nat. Sustain, 1, 441 (2018).
- 108. Renard, D. & Tilman, D. National food production stabilized by crop diversity. Nature 571, 257-260 (2019).
- 109. McDermott, M., Mahanty, S. & Schreckenberg, K. Examining equity: a multidimensional framework for assessing equity in payments for ecosystem services. Environ. Sci. Policy 33. 416-427 (2013).
- 110. Lubchenco, J., Cerny-Chipman, E. B., Reimer, J. N. & Levin, S. A. The right incentives enable ocean sustainability successes and provide hope for the future. Proc. Natl Acad. Sci. USA 113, 14507-14514 (2016).
- Galaz, V., Gars, J., Moberg, F., Nykvist, B. & Repinski, C. Why ecologists should care about financial markets. Trends Ecol. Evol. 30, 571-580 (2015).
- Scholtens, B. Why finance should care about ecology. Trends Ecol. Evol. 32, 500-505 112. (2017)
- 113. Jouffray, J.-B., Crona, B., Wassénius, E., Bebbington, J. & Scholtens, B. Leverage points in the financial sector for seafood sustainability. Sci. Adv. 5, eaax3324 (2019).
- Sumaila, U. R., Lam, V., Le Manach, F., Swartz, W. & Pauly, D. Global fisheries subsidies: an updated estimate. Mar. Policy 69, 189-193 (2016).
- Pe'er, G. et al. EU agricultural reform fails on biodiversity. Science 344, 1090-1092 (2014).
- 116. Central banks and supervisors network for greening the financial system (NGFS). Climate Action in Financial Institutions https://www.mainstreamingclimate.org/ngfs/ (2019).
- 117. Galaz, V. et al. Tax havens and global environmental degradation. Nat. Ecol. Evol. 2, 1352-1357 (2018)
- 118. Miller, D. D. et al. Cutting a lifeline to maritime crime: marine insurance and IUU fishing. Front Ecol Environ 14 357-362 (2016)
- 119. Gardner, T. A. et al. Transparency and sustainability in global commodity supply chains. World Dev 121 163-177 (2019)
- 120. Spink, J. & Moyer, D. C. Defining the public health threat of food fraud. J. Food Sci. 76, R157-R163 (2011).
- 121. Helyar, S. J. et al. Fish product mislabelling: failings of traceability in the production chain and implications for illegal, unreported and unregulated (IUU) fishing. PLoS ONE 9, e98691 (2014).
- 122. Nakamura, K. et al. Seeing slavery in seafood supply chains. Sci. Adv. 4, e1701833 (2018).
- 123. Boström, M., Jönsson, A. M., Lockie, S., Mol, A. P. J. & Oosterveer, P. Sustainable and responsible supply chain governance: challenges and opportunities. J. Clean. Prod. 107,
- 124. Wognum, P. M., Bremmers, H., Trienekens, J. H., Van Der Vorst, J. G. A. J. & Bloemhof, J. M. Systems for sustainability and transparency of food supply chains - current status and challenges. Adv. Eng. Inform. 25, 65-76 (2011).
- 125. Neumann, B. & Unger, S. From voluntary commitments to ocean sustainability. Science **363**. 35-36 (2019).
- 126. Francisco, K. & Swanson, D. The supply chain has no clothes: technology adoption of blockchain for supply chain transparency, Logistics 2, 2 (2018).
- 127. Caro, M. P., Ali, M. S., Vecchio, M. & Giaffreda, R. Blockchain-based traceability in Agri-Food supply chain management: a practical implementation. In 2018 IoT Vertical and Topical Summit on Agriculture-Tuscany 1-4 (2018).
- 128. Hardt, M. J., Flett, K. & Howell, C. J. Current barriers to large-scale interoperability of traceability technology in the seafood sector. J. Food Sci. 82 (S1), A3-A12 (2017).
- 129 Kates, R. W. et al. Sustainability science, Science 292, 641-642 (2001).
- 130. Dauvergne, P. & Lister, J. Big brand sustainability: governance prospects and environmental limits. Glob. Environ. Change 22, 36-45 (2012).
- Barbier, E. B., Burgess, J. C. & Dean, T. J. How to pay for saving biodiversity. Science 360, 486-488 (2018).
- 132. Österblom, H., Jouffray, J.-B., Folke, C. & Rockström, J. Emergence of a global sciencebusiness initiative for ocean stewardship. Proc. Natl Acad. Sci. USA 114, 9038-9043 (2017).
- 133. Arneth, A. et al. Climate Change and Land. An IPCC Special Report on Climate Change, Desertification, Land Degradation, Sustainable Land Management, Food Security, and Greenhouse Gas Fluxes in Terrestrial Ecosystems (IPCC, 2019).
- Westley, F. et al. Tipping toward sustainability: emerging pathways of transformation. Ambio 40, 762-780 (2011).
- 135. Nyborg, K. et al. Social norms as solutions. Science 354, 42-43 (2016).
- 136. Lenton, T. M. & Latour, B. Gaia 2.0. Science 361, 1066-1068 (2018).
- Jørgensen, P. S. et al. Antibiotic and pesticide susceptibility and the Anthropocene operating space. Nat. Sustain. 1, 632-641 (2018).
- 138. Heap, I. & Duke, S. O. Overview of glyphosate-resistant weeds worldwide. Pest Manag. Sci. 74, 1040-1049 (2018).

- 139. Knuth, S. E. Global finance and the land grab: mapping twenty-first century strategies. Can. J. Dev. Stud. 36, 163-178 (2015).
- 140. Clapp, J. & Isakson, S. R. Risky returns: the implications of financialization in the food system. Dev. Change 49, 437-460 (2018).
- Epstein, G. in Financialization and the World Economy (ed. Epstein, G. A.) 3-16 (2005).
- 142. Sullivan, S. Banking Nature? The spectacular financialisation of environmental conservation. Antipode 45, 198-217 (2013).
- 143. Golden, J. S. et al. Making sure the blue economy is green. Nat. Ecol. Evol. 1, 17 (2017).
- 144. First World Bank 'blue bond' approved for Seychelles. SDG Knowledge Hub https://sdg. iisd.org/news/first-world-bank-blue-bond-approved-for-seychelles/ (2017).
- 145. Knott, C. & Neis, B. Privatization, financialization and ocean grabbing in New Brunswick herring fisheries and salmon aquaculture. Mar. Policy 80, 10-18 (2017).
- 146. Benediktsson, K. & Karlsdottir, A. Iceland: crisis and regional development—thanks for all the fish? Eur. Urban Reg. Stud. 18, 228-235 (2011).
- 147. Bodin, Ö. Collaborative environmental governance: achieving collective action in socialecological systems. Science 357, eaan1114 (2017).
- 148. Epstein, G. et al. Institutional fit and the sustainability of social-ecological systems. Curr. Opin, Environ, Sustain, 14, 34-40 (2015).
- 149. Bodin, O., Crona, B., Thyresson, M., Golz, A. L. & Tengö, M. Conservation success as a function of good alignment of social and ecological structures and processes. Conserv. Biol. 28, 1371-1379 (2014).
- 150. Barnes, M. L. et al. Social-ecological alignment and ecological conditions in coral reefs. Nat. Commun. 10, 2039 (2019).
- 151. Gelcich, S. et al. Navigating transformations in governance of Chilean marine coastal resources. Proc. Natl Acad. Sci. USA 107, 16794-16799 (2010).

Acknowledgements We thank R. Cottrell for sharing data on global food production shock frequency. We are grateful to Mistra, the Beijer Foundation, the Erling-Persson Family Foundation and the Swedish Government for providing funding. J.-B.J. and P.S.-J. were supported by the Swedish Research Council Formas (project numbers 2015-743 and

Author contributions M.N. conceived the original idea. M.N. and J.-B.J. developed the idea and led the writing with support from A.V.N. All authors contributed to the development and

Competing interests J.-B.J., B.C., P.S.-J. and C.F. provide scientific support to companies in the seafood sector through the Seafood Business for Ocean Stewardship (SeaBOS) initiative (https://keystonedialogues.earth). The other authors declare no competing interests.

#### Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-

Correspondence and requests for materials should be addressed to M.N. Peer review information Nature thanks Kirsty Lewis, Pablo A. Marquet and the other. anonymous, reviewer(s) for their contribution to the peer review of this work. Reprints and permissions information is available at http://www.nature.com/reprints. Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© Springer Nature Limited 2019

# Genetic strategies for improving crop yields

https://doi.org/10.1038/s41586-019-1679-0

Received: 5 April 2019

Accepted: 16 September 2019

Published online: 6 November 2019

Julia Bailey-Serres<sup>1,2\*</sup>, Jane E. Parker<sup>3</sup>, Elizabeth A. Ainsworth<sup>4,5</sup>, Giles E. D. Oldroyd<sup>6</sup> & Julian I. Schroeder<sup>7,8</sup>\*

The current trajectory for crop yields is insufficient to nourish the world's population by 2050<sup>1</sup>. Greater and more consistent crop production must be achieved against a backdrop of climatic stress that limits yields, owing to shifts in pests and pathogens, precipitation, heat-waves and other weather extremes. Here we consider the potential of plant sciences to address post-Green Revolution challenges in agriculture and explore emerging strategies for enhancing sustainable crop production and resilience in a changing climate. Accelerated crop improvement must leverage naturally evolved traits and transformative engineering driven by mechanistic understanding, to yield the resilient production systems that are needed to ensure future harvests.



The Green Revolution of the 1960s enabled a steep increase in the yields of major staple grain crops (wheat, corn and rice) to address the caloric needs of an increasing global population. This was accomplished through elite variety breeding, hybrid crop development, fertilizer application and advances in management through substantial public investment<sup>2</sup>. The consequent rise in food security benefitted many regions of the world and improved agricultural development (particularly in India and southeast Asia), reducing poverty and malnourishment.

By the 1980s, molecular and transformation technologies propelled the delivery of the first bioengineered genes into plant genomes. Currently, the most widely adopted genetically modified traits are resistance to herbicides and insects in crops with large markets (maize, soybean, cotton and Brassica napus (canola)). Although herbicideand insect-resistance traits greatly lessened soil tillage and insecticide use, respectively, they require careful management to avoid natural selection of resistance in weeds or pests<sup>3,4</sup>. Despite engineered traits with clear benefits to farmers and end-users (including virus-resistant papaya<sup>5</sup>, drought-tolerant corn<sup>6</sup>, rice<sup>7</sup> and bananas<sup>8</sup> fortified with provitamin A, non-browning apples and low-acrylamide potatoes 10, the acceptance of genetically modified traits is equivocal in some countries, and the cultivation of genetically modified crops is largely banned in the European Union.

Future food security will require reducing crop losses due to environmental factors, including climate change, as well as transformative advances that provide major gains in yields. More recent genomic technologies have expedited breeding and trait development for increased environmental resilience and productivity. Genetic diversity is now readily explored at nucleotide-scale precision, using genome-wide association studies and other gene-mapping methods paired with advanced phenotyping systems. The identification of loci that contribute to traits, coupled with molecular-marker-assisted breeding, enables the rapid selection of new genetic combinations in elite varieties. Complementary to breeding approaches, advances in the spatial and temporal regulation of engineered genes and pathways are increasingly accelerated by the targeted editing of genomes using CRISPR-Cas technology. A greater understanding of plant mechanisms that increase yields in variable environments is essential to drive the necessary gains in crop improvement, which can be fuelled by genetic diversity and implemented by genome-scale breeding, finely-tuned gene engineering and more-precise agronomic management practices.

# **Post-Green Revolution challenges**

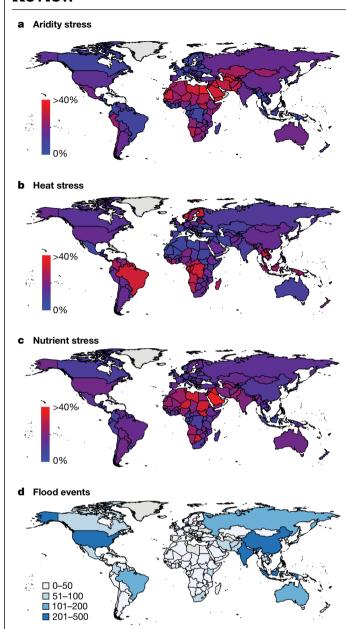
Despite the marked effect of the Green Revolution on food security, there were uneven consequences for human nutrition, the resilience of crops to stress, and the environment<sup>2</sup>. Asian populations benefitted from the increased production of staple grains, and the adoption of irrigation across vast areas<sup>2,11</sup>. The limited rise in food security in sub-Saharan Africa and other impoverished areas can be traced to geographically skewed support and a lack of investment in orphan crops<sup>2</sup>. An unintended consequence has been that fruits and vegetables rich in macronutrients have been displaced by calorie-rich and higher-value grain crops in some areas<sup>2</sup>. Moving forward, an increased production of nutrient-rich vegetable, pulse, tuber and cereal crops, and a broadening of the global reach of agricultural advances, is necessary to achieve food and nutritional security<sup>12</sup>.

# Climatic stress and disease management

The increasing frequency of debilitating heat-waves, droughts, torrential rains and other weather extremes experienced across the globe negatively affects agricultural productivity, and is projected to do so<sup>1,13</sup> (Fig. 1). Climatic constraints can occur independently or together (as with heat and aridity), and in either case reduce the level of productivity that is predicted for a well-managed environment (the yield potential).

1 Center for Plant Cell Biology and Department of Botany and Plant Sciences, University of California Riverside, Riverside, CA, USA. 2 Institute of Environmental Biology, Utrecht University, Utrecht, The Netherlands. 3Department of Plant-Microbe Interactions, Max Planck Institute for Plant Breeding Research and Cluster of Excellence on Plant Sciences (CEPLAS), Cologne, Germany. 4Global Change and Photosynthesis Research Unit, Agricultural Research Service, US Department of Agriculture, University of Illinois at Urbana-Champaign, Urbana, IL, USA. 5Department of Plant Biology, University of Illinois at Urbana-Champaign, Urbana, IL, USA, 6Sainsbury Laboratory, University of Cambridge, Cambridge, UK, 7Cell and Developmental Biology Section, Division of Biological Sciences, University of California San Diego, La Jolla, CA, USA. 8Food and Fuel for the 21st Century, University of California San Diego, La Jolla, CA, USA. \*e-mail: serres@ucr.edu: iischroeder@ucsd.edu

# Review



**Fig. 1** | **Predicted national-scale yield loss for maize, rice, wheat and soybean. a**–**c**, Maps indicate the yield losses caused by aridity stress averaged from 1950–2000 (**a**), heat stress averaged from 1994–2010 (**b**) and nutrient stress in 2009 (**c**). National data for each crop were previously compiled <sup>13</sup>, and are here averaged and re-plotted using the maps package in R<sup>153</sup>. **d**, Number of large flood events from 1985 to 2010 <sup>154</sup> by country.

An unanticipated consequence of the development of high-yield varieties for locations with advanced cultivation practices has been a loss of genetic variation that is associated with resilience to suboptimal environments. It is imperative to breed crops that carry a diversity of resistance genes and/or to plant a diversity of varieties, as this approach minimizes the ability of pathogens to overcome resistance <sup>14</sup>. An increasing occurrence of extreme weather events, together with dire projections of climate change, makes the improvement of crop resilience to environmental (abiotic) and pathogen (biotic) stress of paramount importance for feeding a growing global population.

#### Fertilizer use

The combination of high-yield crop varieties and the widespread use of inorganic fertilizers markedly improved crop production, with clear

benefits in terms of food security<sup>15</sup>. This has translated to excessive anthropogenic release of reactive nitrogen<sup>16</sup> and phosphate<sup>17</sup> into the environment. Inorganic fertilizers have pushed the global nitrogen and phosphorus cycles well beyond their estimated safe operating space<sup>18</sup>, with considerable negative effects on biodiversity, human health and the atmosphere<sup>19</sup>. Their use presents a paradox, as the optimization of plant nutrition stabilizes yields and has helped to reduce expansions in crop area in light of population growth, yet nitrogenous fertilizers contribute substantially to the greenhouse gas emissions that promote climate change<sup>19</sup>.

# **Paths forward**

The agriculture of the next decades must satisfy demands for nutritious food, fibre and animal feed in a highly variable climate, and also mitigate the effects of agriculture on the environment. This is a tall order. Key to addressing the challenge is a deeper understanding of genetic variation and the molecular, cellular and developmental pathways by which plants dynamically respond to and interact with their environment and pathogens, while maintaining growth, efficiency of nutrient use and fitness. New crop varieties ideally will have genetic combinations that alleviate losses from the multiple environmental and pest constraints that are encountered during the crop lifecycle in a farmer's field. An important emerging and non-trivial goal is to optimize the efficiency of photosynthesis, water and nutrient use, including the fostering of beneficial interactions between plants and microorganisms that can promote nutrient acquisition. The integration of mechanistic understanding, genetic variation and genome-scale breeding towards technological solutions will be essential. Here we review advances and emerging directions within the plant sciences that may bolster yield-defining traits and resilience (Fig. 2, Box 1).

## Protection from new and re-emerging diseases

The reliance on major pathogen-resistance genes bred into crop monocultures provides short-term protection against diseases, as seen in the boom-and-bust cycles of resistance over the past century and in the spread of new diseases across continents. A more complete molecular-genetic framework now exists for general and specific resistance to microbial pathogens mediated by a two-layered receptor signalling system (Fig. 2a). At the plant cell surface, families of pattern-recognition receptors register the presence of microorganisms<sup>20</sup>. Inside host cells, large panels of nucleotide-binding domain leucine-rich-repeat receptors (NLRs) detect activities of invasive pathogenic strains<sup>21</sup>. Advances in elucidating receptor-pathogen recognition and activation mechanisms at a protein-structural level provide strategies towards the rational design of receptor proteins that are tailored to intercept broader or alternative disease agents <sup>21,22</sup>. Newly engineered resistance traits can then be transferred to elite varieties of crops to confer resistance against modern diseases. There have been notable successes in the inter-family transgenic transfer of a pattern-recognition-receptor gene to potato, tomato, Medicago, wheat and rice<sup>20,23</sup>, indicating that surface receptors that are restricted to particular plant lineages can confer immunity in unrelated species. Transfer of the wheat *Pm3e* resistance gene against powdery mildew to a susceptible wheat variety has produced effective mildew resistance in field trials<sup>24</sup>. The engineering of pathogen-induced translational control of a key *Arabidopsis* immunity component in rice<sup>25</sup> has provided promising disease-resistance benefits in initial crop field trials, apparently without a yield penalty. The incorporation of new surface- and intracellular-receptor recognition and signal transduction modules into crops is also on the horizon, building on knowledge of receptor functional partnerships and resistance network archi $tectures ^{21,26}. \, Success \, in \, this \, area-especially \, as \, climates \, change-will \, also \, climates \, change-will \, consider a constant of the consta$ require tight immune-receptor control, which can require co-evolved

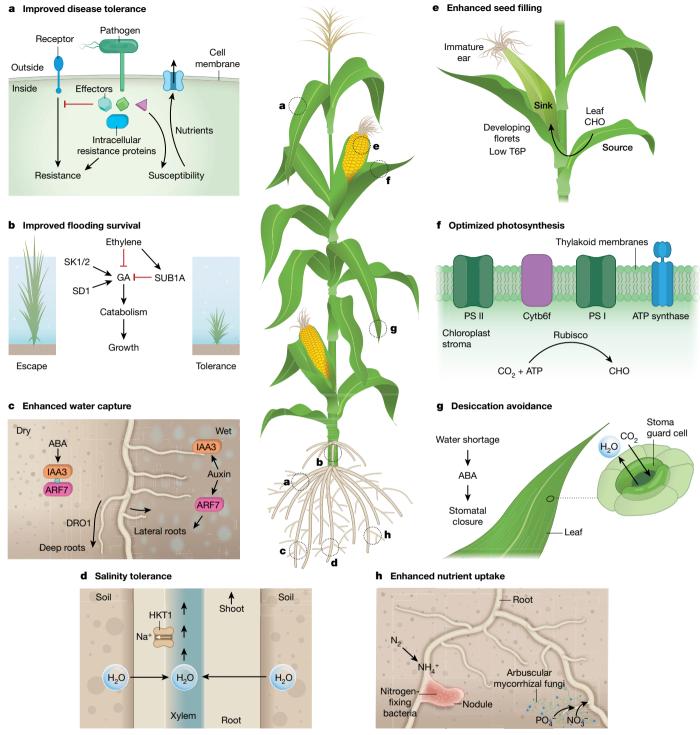


Fig. 2 | Paths to increased crop yield in suboptimal environments. Overview of traits that provide increased resilience and yield in variable environments. a, Pathogen recognition by cell-surface and intracellular receptors (resistance proteins). Manipulation of host cells by pathogen-secreted effectors to promote infection can be recognized by resistance proteins and converted to  $disease \ resistance. \ \textbf{b}, Flooding \ survival \ via \ opposing \ regulation \ of \ gibberellin$ (GA). Semidwarf 1 (SD1), Snorkel 1 and Snorkel 2 (SK1/2) confer escape by accelerated elongation growth. Submergence 1A (SUB1A) confers tolerance by quiescence of growth. c, Root growth towards moisture involves transcriptional regulators (indol-3-acetic acid inhibitor protein 3 (IAA3) and

 $auxin\,response\,factor\,(ARF7)), and\,is\,regulated\,by\,the\,hormones\,ABA\,and$ auxin. d, HKT1 (high-affinity K transporter sub-family 1) mediates sodium (Na<sup>+</sup>) exclusion from leaves. e, In developing seed tissues, catabolism of T6P aids the movement of photo-assimilate carbohydrate (CHO) from leaves to sinks in developing florets. f, Optimizing photosynthetic light harvesting and  $CO_2 fix at ion \, by \, altering \, photosynthetic \, protein \, abundance \, and \, minimizing \,$ photorespiration. PS, photosystem. g, Dynamic control of stomatal aperture by pairs of epidermal guard cells lessens desiccation. h, Symbiotic plantmicroorganism interactions facilitate the uptake of essential nutrients. NH<sub>4</sub><sup>+</sup>, ammonium; PO<sub>4</sub>-, phosphate; NO<sub>3</sub>-, nitrate.

# Box 1

# Yield-defining traits and opportunities for crop improvements

This Review discusses several traits that are essential for crop performance, including the genetic variation and plasticity that are relevant for improvement (left) and advanced and emerging approaches for addressing trait improvements (right). Stress resilience coupled with high yields is aided by hardwired traits and temporal responses to a dynamic environment. Opportunities for improvement include capturing natural genetic variation, the functional characterization of genes, the manipulation of endogenous or transferred genes with appropriate regulatory control, the development of low-cost and safe small molecules that can be delivered to plants before stress or during recovery, and improved plant health through interactions with symbiotic microorganisms.

#### **Yield-defining traits**

## Shoot traits and plasticity

Inflorescence architecture and fertility

Shoot-to-root biomass

Photosynthesis

Stomatal movement and density regulation

Assimilate loading and partitioning

Senescence timina

#### Root traits and plasticity

Architecture and anatomy

Growth dynamics

Nutrient acquisition and use efficiency

Microbial interactions

#### Stress resilience

Drought, salinity, flooding and extreme temperatures (abiotic)

Pests and pathogens (biotic)

Tempered response to minimize growth penalty

#### **Opportunities**

# Natural genetic variation

Stress resilience and recovery mechanisms

Trait pyramiding

#### Gene engineering and editing

Spatial, temporal and inducible control of genes and networks

Improving protein function, targeting and turnover

Enhancing metabolite pathway and flux

Introducing synthetic traits

#### Beneficial soil and leaf microbiome

Seeding and supplementation

Attraction of beneficial microorganisms

#### Small-molecule delivery

Response activation

Metabolic regulation

#### Sensor use for crop management

Cellular, organ, canopy and remote

factors <sup>27,28</sup>, because modified receptors without requisite constraints are often mis-activated and cause necrosis and reduced plant health. Knowledge in this arena may also lead to strategies that lessen food spoilage by pathogens after harvesting.

# Harnessing resistance from diverse germplasm

With increased access to diverse genetic variation found in crops and natural populations of wild relatives<sup>14</sup>, the door is open to recovering disease-resistance traits: many of these are encoded by genes for pattern recognition and NLRs that were lost during domestication, or that have evolved independently in different plant lineages<sup>14,29</sup>. Advances in genome sequencing and assembly technologies, coupled with new methods for capturing near-complete immune-receptor gene panels from complex genomes, hold promise for attaining sustainable disease resistance<sup>30–32</sup>. Merging these approaches with genome-wide association studies taps into immune-receptor gene variants that have adapted to local environments and pathogen populations to help to increase the resilience of future crops. The stacking (or 'pyramiding') of several resistance genes with different recognition spectra and environmental optima into a single background is now a credible strategy for achieving more durable disease resistance. Nevertheless, assembling appropriate gene combinations in elite varieties of crops remains a challenge.

Investigations of plant genomes within and across species provide insight into the evolutionary forces that have shaped the architecture and function of genes related to immunity. This will aid the design of new resistance traits<sup>33</sup>. The isolation and characterization of genes associated with disease susceptibility in the host has also gained prominence<sup>34</sup>. The proactive deployment of modified susceptibility genes in crops will become possible as geographical sampling of pathogen genomes and populations increases<sup>35,36</sup>. A recessive barley *mildew resistance locus o (mlo)* mutant that breeders have successfully used for 75 years against powdery mildew disease has been engineered into hexaploid wheat using mutagenesis by transcription-activator-like

effector nuclease, or by combining mutations selected in three wheat MLO loci  $^{37,38}$ . Thus, the characterization and manipulation of host–pathogen infection processes for the establishment of disease can generate novel resistance mechanisms in crops that are not necessarily found in natural populations. As a cautionary note, some newly engineered crop lines have displayed unexpected phenotypes and vulnerabilities to disease  $^{39}$ , which underscores the need for rigorous performance testing of new material in field settings over multiple seasons.

## Pathogen resistance in a shifting climate

Given the current trends in climate, attaining durable resistance in high-vield crops will require a greater knowledge of pathogen population dynamics and plant host responses to temperature. The alarming spread of devastating disease agents—such as the bacterium Xylella fastidiosa that attacks olives and woody crops in southern Europe, or the ug99 strain of stem rust fungus Puccinia graminis f. sp. tritici that affects wheat across parts of Africa and Asia-is attributed in part to warmer climates, and presents a complex biogeographical and epidemiological problem<sup>40,41</sup>. Moreover, surface- and intracellular-immunity systems appear to respond differently to changes in temperature, owing to effects on the microorganisms and hosts that are not completely understood<sup>42,43</sup>. Although immunity mediated by intracellular receptors tends to be less efficient as temperatures increase, some resistance genes confer protection at higher temperatures 44,45. Although these findings are reassuring, they highlight the need for more genotype  $\times$  environment studies in crops to stabilize resistance in increasingly precarious climates.

### Resilience to abiotic stress

Abiotic stresses associated with climate change that destabilize yields include flooding, drought, soil salinity and extreme temperatures (Fig. 1). Resilience mechanisms have been mobilized for crop improvement through the identification of genes that are associated with key

traits and signal transduction pathways, followed by breeding or engineering<sup>46,47</sup> (Fig. 2b-f). Attaining resilience without affecting overall vield is a considerable challenge.

#### **Flooding**

Floods regularly limit yields<sup>48</sup>. Rice is exceptionally resilient against flooding, yet over 30% of the acreage cultivated with rice suffers yield loss owing to plant submergence<sup>49</sup>. SUBMERGENCE1 (SUB1), identified in a flooding-tolerant landrace of rice, encodes a cluster of genes for ethylene-responsive transcription factors, including the submergenceactivated SUB1A-1<sup>50</sup>. The SUB1A transcription factor curbs the activation of genes that promote the breakdown of leaf sugars and starch (photoassimilate) that would otherwise fuel growth to escape an inundation<sup>51</sup> (Fig. 2b). The introduction of SUB1A-1, through marker-assisted breeding, into high-yield varieties now provides an additional week or more of submergence tolerance-without compromising yields under nonsubmergence conditions<sup>52</sup>.

The submergence tolerance by growth quiescence provided by SUB1A-1 and SUB1A-1 are the submergence tolerance by growth quiescence provided by SUB1A-1 are the submergence tolerance by SUB1A-1 are the submergence by SUB1A-1 are the1 contrasts with the accelerated underwater elongation of the shoots of varieties of crops that are adapted to progressive seasonal floods in delta regions. Deepwater varieties invest photo-assimilate into the extension of submerged stem internodes (Fig. 2b). This requires the SNORKEL 1 (SK1) and SNORKEL 2 (SK2) genes that encode transcription factors that are similar to SUB1A53, as well as biosynthesis of the hormone gibberellin. Gibberellin biosynthesis involves a functional allele of SEMIDWARF1 (SD1)<sup>54</sup>, the gene that—when mutated—determines the short stature of Green Revolution rice55. This knowledge can improve yields in low-lying areas that affected by climate change.

The alleles of SUB1A, SK1, SK2 and SD1 that are key to flooding survival are found in wild *Oryza* species<sup>56</sup>, which indicates that they arose in ancestral populations in flooded ecosystems. Evolution has modified the same growth-response network involving the hormones ethylene and gibberellin to achieve submergence tolerance or escape in diverse species of wetland plants<sup>48</sup>. Pathways to improved flooding tolerance include manipulation of root traits associated with waterlogging tolerance that involve a conserved mechanism<sup>48</sup> and the oxygen-dependent turnover of SUB1A-1-like transcription factors, accomplished in several species<sup>57–60</sup>. There are other opportunities to protect yields in wet climates. Torrential rain and hail can cause yield losses of 50% or more, owing to premature pod shattering in oil crops. The identification of genes that control pod shattering in Arabidopsis<sup>61</sup> enabled the genetargeted molecular breeding of optimized pod-shattering properties in canola that is now increasingly planted by farmers.

## **Drought**

Drought and other dehydration or osmotic stresses (salinity and cold) stimulate the production of the hormone abscisic acid (ABA) in plants. Although the mechanisms of the initial sensing of osmotic stress and signalling in response to osmotic stress remain poorly understood, the elucidation of the ABA receptor and signal transduction mechanisms<sup>62,63</sup> has exposed new avenues for the enhancement of dehydration tolerance. This includes ABA receptor overexpression<sup>64,65</sup> and engineering to  $respond to exogenously sprayed small molecules {}^{66,67}, the overexpression$  $of signal transduction components {}^{68-70} or the drought-driven repression\\$ of negative regulators of ABA signal transduction<sup>69,71</sup>.

ABA closes the adjustable stomatal pores on the leaf surface that allow gas exchange and thus reduces the water lost from plants during drought, but this response can be weak in crop varieties <sup>69,72</sup>. ABA also helps to regulate root growth in response to water availability, including inhibition of lateral root growth and enhancement of primary and secondary root growth. This developmental reprogramming allows roots to seek water. The DEEPER ROOTING (DRO1) gene of rice provides a deep root architecture in paddy fields, which bolsters yields under water-limited conditions<sup>73-75</sup> (Fig. 2c). The identification of the major loci that control root traits associated with drought resilience has proven challenging owing to their quantitative nature and low heritability, which requires sophisticated belowground phenotyping and analytical methods<sup>76</sup>. Yet roots grow laterally towards moisture in soil<sup>77</sup>. New roots that access moisture emerge only on the side of a root that is moist, as a modification of a key auxin-response transcription factor on the dry side of a root impedes the developmental program<sup>78</sup> (Fig. 2c). Knowledge such as this can inform strategies for the advanced breeding and engineering of improved resilience to drought, which continues to limit yields<sup>79,80</sup>.

# **Salinity**

Irrigation substantially expands growing seasons and increases crop yields in many regions. Salt (sodium chloride) gradually accumulates in irrigated soils and is toxic to most crops; sodium accumulation is particularly detrimental in leaves. Approximately 40% of irrigated lands worldwide are affected by increased salt levels, and expansion of soil salinization is a major threat to crop performance<sup>47</sup>.

Plants encode a sodium-transporter gene sub-family named HKT181 (high-affinity K+transporter) that provides protection from the over $accumulation of so dium in leaves {}^{82} (Fig. 2d). HKT1 mediates the removal\\$ of sodium, mainly in roots, from the xylem<sup>83,84</sup>, the vascular conduits that transport water and nutrients from roots to leaves. Major quantitative trait loci that enhance salt tolerance in wheat, wheat relatives and rice possess distinct *HKT1* alleles <sup>47,83,85</sup>. This knowledge has enabled the marker-assisted breeding of wheat with a higher salt tolerance, resulting in a yield improvement of 25% under salinity stress in field trials<sup>85</sup>. Beneficial alleles of *HKT1* may enhance salinity tolerance in other species, as has been shown in rice83. It will be necessary to combine HKTs with other strategies to further boost salinity resistance as land salinization continues to rise. The effects of salinity on root development also need to be factored into intervention strategies 86,87. Natural variation in transporter genes and their regulation has also provided field-tested solutions for other toxic elements, including aluminium<sup>88,89</sup> and boron<sup>90</sup>.

# **Extreme temperatures**

Higher atmospheric levels of CO<sub>2</sub> and other greenhouse gases are predicted to increase the frequency and duration of heat-waves<sup>91</sup>, which will lead to losses in crop yield—especially in arid regions<sup>92</sup>. Sensitivity to extreme temperatures varies during the plant lifecycle and across species. Low temperatures influence the germination, establishment. growth and viability of crops, except for those with temporal chilling or freezing resilience (such as winter wheat). Genetic variation in key transcriptional regulators of cold resilience is leveraged in breeding of several grain crops<sup>46</sup>. By contrast, warm temperatures promote growth until a threshold is reached above which yields precipitously diminish, especially when soil moisture is low or humidity is high 93,94. Sensitivity to temperature extremes is height ened during reproduction, when it reducesmale fertility and seed quality95. This presents a daunting challenge as protective responses are typically accompanied by reduced yield. Heat stress is an expanding threat in tropical regions, because, at high humidity, plants are less able to cool their leaves by transpiration via stomatal pores that control the trade-off between CO<sub>2</sub> intake and water loss <sup>96</sup>. There is an urgent need for research and for ensuing genetic and engineered solutions that preserve crop productivity at increased temperatures (Box 1).

## Metabolic control of resilience and yield

Breeding or engineering plants for a high yield potential in varied and variable environments is a potential solution for capturing effective resilience. Plants typically dampen growth and accelerate reproductive development as a consequence of stress. Yield maintenance under

# Review

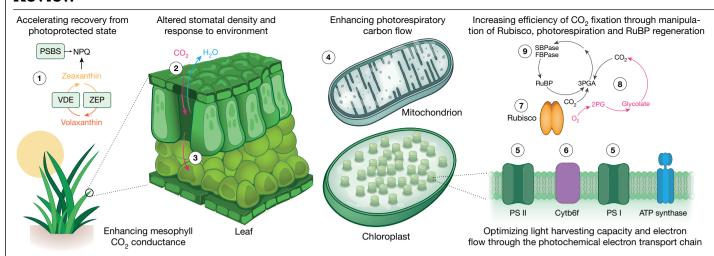


Fig. 3 | Targets for improving the efficiency of photosynthesis and primary carbon metabolism that have experimental support for success. Transgenic manipulations of photosynthetic metabolism that lead to improved photosynthetic efficiency include (1) improving photosynthesis in a dynamic light environment by accelerating recovery from a photoprotected state, by overexpressing enzymes (such as photosystem II subunit S (PSBS) and VDE) that are involved in non-photochemical quenching (NPQ) (the dissipation of excess excitation energy as heat)116; (2) altering the CO2 response of stomata or the

density of stomata on the leaf surface to increase the efficiency of water use<sup>120,122,123</sup>; (3) increasing the capacity for mesophyll conductance of CO<sub>2</sub><sup>105</sup>; (4) improving the energy efficiency of carbon metabolism by altering mitochondrial enzymes<sup>155</sup>; (5) optimizing investment in light collection<sup>105</sup>; (6) increasing electron flow through the photosynthetic electron transport chain<sup>110</sup>; (7) altering Rubisco properties and activation to increase CO<sub>2</sub> assimilation<sup>113,156</sup>; (8) bypassing photorespiration<sup>117</sup>; and (9) increasing the efficiency of ribulose 1,5-bisphosphate (RuBP) regeneration<sup>115</sup>.

moderate drought was significantly improved in AQUAmax corn hybrids produced by selective and marker-assisted breeding<sup>97</sup>. The underlying genetic variants and mechanisms that enable these lines to conserve soil moisture and delay the accumulation of biomass until grain filling remain to be characterized. Higher yields under well-watered conditions as well as under moderate drought at the time of flowering was achieved in corn that expresses a metabolic enzyme that converts the low-abundance metabolite trehalose-6-phosphate (T6P) to trehalose in the phloem companion cells at the base of the ear and developing florets <sup>98</sup> (Fig. 2e). This cell-specific modulation of T6P augments the mobilization of photosynthate to the unfertilized floret, and prolongs the photosynthetic activity of leaves during grain filling. The spatial modulation of T6P levels also regulates the draw of seed reserves into young elongating shoots of rice, particularly when dry seeds are sown directly into a flooded paddy<sup>99</sup>. The application of plant-permeable T6P analogues to wheat leaves increased seed filling and improved recovery from drought<sup>100</sup>. These examples illustrate the critical integration of metabolism and stress resilience to improve crops that can be provided by genetic variation and engineering.

# Optimization of photosynthesis for yield

Modern crops are highly efficient at rapidly spreading their leaf canopies to maximize light interception, and at partitioning carbon and nutrients into seeds. However, crops are not as efficient at converting absorbed light energy into sugars through the process of photosynthesis<sup>101</sup>. This may be because the proteins and enzymes that mediate photosynthesis evolved in a low-light marine environment, which was very different from modern agronomic and atmospheric conditions 102. However, the conservation of chloroplast transmembrane proteins that collect light energy and participate in electron transfer reactions within the chloroplast, along with the conservation of enzymes involved in carbon fixation, reduction and regeneration across plant species 103 (Fig. 2f), has  $aided the modelling of photosynthesis ^{104} and identification of numerous \\$ targets for increasing its efficiency<sup>101</sup> (Fig. 3). Theoretical targets include  $expanding and optimizing light capture by the leaf can opy {\it 105-107}, inducing$ a more rapid relaxation of non-photochemical quenching at photosystem II<sup>108</sup>, increasing the carboxylation capacity of the Rubisco enzyme as well as minimizing oxygenation and photorespiration <sup>109</sup>, enhancing the regenerative capacity of the carbon reduction cycle<sup>101</sup>, optimizing the  $electron transport chain ^{110}, converting crops from C_3 to C_4 metabolism ^{111},\\$ and adding components of cyanobacterial or algal systems to pump CO<sub>2</sub> or compartmentalize Rubisco<sup>101</sup>. Improving photosynthetic efficiency is neither a new nor a universally accepted idea. Some have argued that the selection pressures endured by photosynthesis render it unamenable to improvement<sup>112</sup>. Despite decades of research, the challenge of engineering Rubisco for improved specificity and carboxylation rate remains unmet<sup>113</sup>. However, some recent successes in engineering photosynthetic enzymes and introducing novel pathways into chloroplasts may lead to substantial gains in crop performance, as outlined below.

Maize photosynthesis and fresh weight were enhanced by overexpressing the small and large subunits of Rubisco, together with an assembly chaperone protein<sup>114</sup>. In wheat, the overexpression of sedoheptulose-1,7-biphosphatase showed increased photosynthesis, and resulted in increased plant and grain biomass<sup>115</sup>. These genetic modifications to key crops are promising; their ultimate potential can be tested by incorporating the changes into elite varieties, and evaluation in the field. Photosynthetic manipulations also show promise in the field in the model plant tobacco. Re-engineering the expression of enzymes that control the induction, relaxation and amplitude of non-photochemical quenching successfully enhanced photosynthesis during natural light transitions, which resulted in 14-20% greater vegetative biomass in the field<sup>116</sup>. Even greater gains were observed by inserting enzymes involved in glycolate metabolism into chloroplasts to reduce photorespiration<sup>117</sup>. Coupling this with reduced expression of a glycolate and glycerate transporter, to minimize glycolate flux out of the chloroplast, raised vegetative biomass by 40% under field conditions<sup>117</sup>. These studies represent fundamental breakthroughs in understanding and engineering photosynthesis, which can now move from the proof-of-concept stage to expanded field testing.

# Rising atmospheric CO<sub>2</sub> and plant water loss

Crops lose between 100 and >400 water molecules through stomatal pores in leaves for every carbon atom that is fixed by photosynthesis, highlighting a fundamental trade-off between carbohydrate

production and water use. Increases in CO<sub>2</sub> concentrations inside leaves cause a reduction in the size of stomatal-pore apertures<sup>118</sup>. The continuing rise in atmospheric CO<sub>2</sub> is increasingly narrowing stomatal pores, which can enhance the efficiency of water use by crops. However, many crops have weak or non-optimal stomatal CO<sub>2</sub> responses. Advances have been made in understanding the signal transduction pathways that regulate water loss in response to CO<sub>2</sub><sup>118</sup>. including that the stomatal CO<sub>2</sub> response requires amplification bybut also includes unique components upstream of and parallel to—the ABA response pathway in guard cells<sup>119</sup> (Fig. 2g). The upregulation of the stomatal CO<sub>2</sub> response by guard-cell-targeted overexpression of carbonic anhydrases increased instantaneous water-use efficiency by about 44% in *Arabidopsis*, without a reduction in photosynthetic assimilation rates at ambient CO<sub>2</sub><sup>120</sup>. On the other hand, C<sub>2</sub> crops growing in nutrient-rich and water-sufficient humid regions could benefit from a weaker CO<sub>2</sub>-induced stomatal closing response, which may enhance growth owing to CO<sub>2</sub> 'fertilization' in an atmosphere with an increased concentration of CO<sub>2</sub><sup>121</sup>. A complete understanding of the CO<sub>2</sub> response pathway is needed to optimize and test water-use efficiency and gas-exchange strategies in the field.

Successful transgenic modifications in barley and rice have shown that reducing the density of stomata improves plant performance under water-restricted conditions 122,123. Overexpression of the chloroplast photo system II subunit S protein in to bacco was reported to lower stomatal conductance, and increased the efficiency of water use by field-grown plants<sup>124</sup>. The effective manipulation of stomatal function will require the discovery of the primary CO<sub>2</sub> and/or bicarbonate sensors that control the stomatal CO<sub>2</sub> response, as well as harnessing natural genetic variation in stomatal properties that could improve trade-offs between carbon gain and water loss in a world with high levels of atmospheric  $CO_2^{125}$ .

# Technologies to reduce fertilizer use

Yields of crops are heavily dependent on sufficient nutrition (in particular, nitrogen and phosphorus) that is currently provided primarily through the application of inorganic fertilizers. In smallholder farming systems, crop productivity is limited by the availability of these nutrients15. Substantial advances have been made in understanding the mechanisms of nutrient uptake, transport and use in plants, with the aim of improving sufficiency<sup>126</sup> (Fig. 2h). Balancing the use of photo-assimilate with nutrient uptake is critical for optimizing vields. The mutations that confer stem shortening in cereals, which facilitated the Green Revolution, brought with them unintended inefficiencies in nitrogen use that can be compensated for by changing the balance of transcription factors that control growth and nutrient use<sup>127</sup>. Breeding can also contribute to reducing nutrient imbalances through the optimization of rooting systems, nutrient transport activity and partitioning128.

In natural ecosystems, plants frequently engage with beneficial microorganisms that facilitate the uptake of limiting nutrients such as nitrogen and phosphate<sup>129</sup>. In agriculture, these beneficial associations are often dampened by supplied fertilizers, because plants suppress their interaction with symbionts when they perceive ample nutrients. Most plant species associate with arbuscular mycorrhizal fungi that greatly expand the root-surface area for nutrient uptake, and which actively mine immobilized phosphates from the soil 130,131. Bringing associations with arbuscular mycorrhizal fungi more effectively into annual cropping systems with moderate fertilizer use could improve nutrient capture, and increase sustainability-particularly if the phosphate suppression of mycorrhization could be overridden. However, applications of strigolactones (the plant-derived lowphosphate signals to microorganisms) have so far been insufficient to override the suppression of mycorrhization<sup>132</sup> and more research is therefore needed to obtain benefits from mycorrhizal associations in agriculture.

#### **Engineering the nitrogen symbiosis**

Some plants are colonized intracellularly by nitrogen-fixing bacteria that can deliver the complete nitrogen needs of the host plant. Associations such as this are limited to a select group of species, which presents an opportunity to radically improve nitrogen availability for cereal crops if the symbiosis trait can be transferred. Multiple avenues are being explored to achieve this-from equipping plants to associate with nitrogen-fixing bacteria to the transfer of the enzyme nitrogenase, which is responsible for nitrogen fixation. Studies of these processes in their native context have provided an understanding that was absent 30 years ago, when such approaches were first broached. Evolutionary, genomic and mechanistic studies suggest that relatively few genetic components might be needed to confer nitrogen-fixation capabilities. In the case of transferring nitrogenase to plants, the restriction of genetic components required was achieved by concatermerizing bacterial genetic units to create a minimal set of three genes that are necessary for the transfer of nitrogen fixation<sup>133</sup>. Moreover, some components of nitrogenase can be stably expressed in yeast and plants<sup>134</sup>.

The evolution of the nitrogen-fixing symbiosis in legumes used many  $components\,that\,function\,in\,associations\,with\,arbuscular\,my corrhizal$ fungi<sup>129</sup>, which means that cereals possess some of the necessary building blocks and have the potential to streamline engineering efforts to transfer the nitrogen-fixing symbiosis. Recent phylogenomic approaches indicate that very minimal gene reduction (between two and seven genes) is associated with the loss of nitrogen fixation 135,136, suggesting that a small set of genes could convert a species that associates with arbuscular mycorrhizal fungi into one that can also form nitrogen-fixing symbiosis. This considerable engineering challenge will require precise transcriptional and post-translational regulation of multiple heterologous genes in cereals. Additional challenges are the few well-characterized promoters for gene regulation in cereal roots, and bottlenecks associated with transformation of cereals that limit the scale of throughput required to test the engineering iterations that will be necessary to achieve nitrogen fixation.

# Benefits of plant-associated microorganisms

The environment around and within plant roots includes microbial communities. These can be relatively restricted<sup>137</sup> or dynamic<sup>138</sup>, and responsive to nutrient status<sup>139</sup>. Such communities or community members have the potential to protect plants against pathogen infections 140,141 and, to some extent, drought 142,143. A greater understanding of the mechanisms and the environmental conditions, including climate effects<sup>144</sup>, that control plant-microorganism assembly and activities may enable the engineering of microbial communities to optimize crop performance, particularly with microorganisms that are engineered using synthetic biology approaches. Current research indicates that some fungal species benefit host plants by enhancing phosphate uptake<sup>145</sup>, and within the diversity of cereal crops are lines that can host active communities of nitrogen-fixers<sup>146</sup>. The manipulation of microbial associations to improve crop resilience to environmental stresses is an area of intense research.

# **Prospectus on resilient crops**

Research advances have provided innovative opportunities and technologies across the plant sciences, which can furnish solutions for addressing future food security (Box 1). The strategies described here for enhancing the resilience and sustainability of crops will only be realistic if they are part of an integrated approach to agriculture that is developed collaboratively with agronomists, engineers and farmers<sup>147</sup>. A critical challenge is the time from research discovery to true and widespread implementation in agriculture. Some high-impact breeding and genetically modified traits (for example, pest resistance mediated by individual Bacillus thuringiensis Cry proteins) have spread relatively rapidly. However, even in cases that involve breeding into

# Review

diverse varieties, the time from initial discovery and development to broad use has often exceeded ten years<sup>79</sup>. Regulatory processes and intellectual property hurdles associated with technology can lead to additional delays in implementation. The robust assessment of varieties in variable field environments is essential to timely adoption. In the case of submergence-tolerant SUB1 rice varieties<sup>50</sup>, cooperation between scientists, breeders and farm advisers helped to achieve farmer acceptance and governmental certification within three years of gene characterization. The visible yield advantage of SUB1 varieties after flooding, and their lack of differences with the varieties they replaced, was key to their adoption. This success contrasts with the failure to provide many farmers in climate-vulnerable areas with the services of plant breeders to mobilize genetic variation for crop improvement <sup>148</sup> and with selected or engineered genotypes that did not translate to the field<sup>46</sup>. Complementary approaches and technologies may provide viable opportunities (such as high-protein, salt-tolerant algae that require limited freshwater)—although new infrastructure, energy inputs and engineering solutions will be needed<sup>149</sup>.

Valuable genetic diversity for increasing crop resilience resides in cultivated landraces, heirloom varieties and the wild relatives of crops. Seed banks curate and distribute crop germplasm; the Crop Trust (https://www.croptrust.org) is one of the leading efforts to collect, conserve and use the approximately 50,000 species of wild relatives of crops<sup>150</sup>. These seed banks distribute germplasm that can be tapped for adaptations to abiotic and biotic stresses, but greater investment in high-throughput genotyping and phenotyping is needed to accelerate mapping, the identification of genes and mechanisms, and downstream breeding151

Addressing yield loss due to climate change, salinity and (re)emerging diseases, weeds, parasitic plants and pests requires innovative technologies and proactive responses, not unlike the development of vaccines and innovations in modern medicine. The integration of genetic resources and transformative technologies, from genome editing to synthetic biology, are necessary to capture traits that increase global food security and reduce the effects of agriculture on the environment. An early failure of plant biotechnologists was in the lack of effective engagement with environmentalists, farmers and consumers on questions of health and safety, despite strict governmental procedures for the validation, release and monitoring of genetically modified crops. It is critical that the specific method used for crop improvement does not stymie the implementation of safe and effective solutions. Non-politicized regulatory systems are essential for scientific advances to scale to farmers within the timeframe needed.

The current timeline for increasing the resilience and sustainability of crops is too long. Crop varieties with new combinations or variants of disease-resistance genes are in preparation for use against newly emerged virulent pathogens. Advances in sequencing and the early detection of invasive pathogenic strains should enable better monitoring of disease, and therefore knowledge regarding where to deploy particular crop genotypes. The horizon for tailored panels of appropriately controlled genes that impart functional immunity in a commercial crop is years away. The most rapid translation to the field will be for small suites of genes from existing crop germplasm. For challenges that are difficult to overcome (such as resilience to heat and aridity during plant sexual reproduction98), disruptive advances such as the asexual propagation of seeds<sup>152</sup> could lessen yield loss due to male infertility. Success in the engineering of improved photosynthesis, nutrient use and beneficial plant-microorganism interactions requires intensive investment, but could result in the gains needed.

The plant sciences have a critical role in meeting the food and fibre challenges of the future. Timely investments and research at many levels and collaborative efforts are paramount to deploying resilience mechanisms and improving the sustainability, yields and nutritional value of our crops.

- Ray, D. K., Mueller, N. D., West, P. C. & Foley, J. A. Yield trends are insufficient to double global crop production by 2050. PLoS ONE 8, e66428 (2013)
- Pingali, P. L. Green revolution: impacts, limits, and the path ahead. Proc. Natl Acad. Sci. USA 109, 12302-12308 (2012).
- Duke, S. O. Perspectives on transgenic, herbicide-resistant crops in the United States almost 20 years after introduction. Pest Manag. Sci. 71, 652-657 (2015)
- Tabashnik, B. E., Brévault, T. & Carrière, Y. Insect resistance to Bt crops: lessons from the first billion acres. Nat. Biotechnol. 31, 510-521 (2013).
- Fitch, M. M. M., Manshardt, R. M., Gonsalves, D., Slightom, J. L. & Sanford, J. C. Virus resistant papaya plants derived from tissues bombarded with the coat protein gene of papaya ringspot virus. Bio/Technology 10, 1466-1472 (1992).
- Castiglioni, P. et al. Bacterial RNA chaperones confer abiotic stress tolerance in plants and improved grain yield in maize under water-limited conditions. Plant Physiol. 147, 446-455 (2008)
- Potrykus, I. From the concept of totipotency to biofortified cereals. Annu. Rev. Plant Biol. 66. 1-22 (2015).
- Paul, J.-Y. et al. Golden bananas in the field: elevated fruit pro-vitamin A from the expression of a single banana transgene, Plant Biotechnol, J. 15, 520-532 (2017).
- Murata, M. et al. A transgenic apple callus showing reduced polyphenol oxidase activity and lower browning potential. Biosci. Biotechnol. Biochem. 65, 383-388 (2001).
- Rommens, C. M., Yan, H., Swords, K., Richael, C. & Ye, J. Low-acrylamide French fries and potato chips, Plant Biotechnol, J. 6, 843-853 (2008).
- FAO. How to Feed the World in 2050, http://www.fao.org/fileadmin/templates/wsfs/docs/ expert\_paper/How\_to\_Feed\_the\_World\_in\_2050.pdf (FAO, 2009).
- 12. Herrero, M. et al. Farming and the geography of nutrient production for human use: a transdisciplinary analysis. Lancet Planet. Health 1, e33-e42 (2017).
- Mills, G. et al. Closing the global ozone yield gap: quantification and cobenefits for multistress tolerance. Glob. Chang. Biol. 24, 4869-4893 (2018).
- Nelson, R., Wiesner-Hanks, T., Wisser, R. & Balint-Kurti, P. Navigating complexity to breed disease-resistant crops. Nat. Rev. Genet. 19, 21-33 (2018).
- Mueller, N. D. et al. Closing yield gaps through nutrient and water management. Nature 490, 254-257 (2012).
- Fowler, D. et al. The global nitrogen cycle in the twenty-first century. Phil. Trans. R. Soc. Lond, B 368, 20130164 (2013).
- Qiu, J. Phosphate fertilizer warning for China. Nature News https://doi.org/10.1038/ news.2010.498 (2010).
- Steffen, W. et al. Planetary boundaries: guiding human development on a changing planet. Science 347, 1259855 (2015).
- Fowler, D. et al. Effects of global change during the 21st century on the nitrogen cycle. Atmos. Chem. Phys. 15, 13849-13893 (2015).
- 20. Boutrot, F. & Zipfel, C. Function, discovery, and exploitation of plant pattern recognition receptors for broad-spectrum disease resistance, Annu. Rev. Phytopathol. **55**, 257-286 (2017).
- Monteiro F & Nishimura M T Structural functional and genomic diversity of plant NLR proteins: an evolved resource for rational engineering of plant immunity. Annu. Rev. Phytopathol 56 243-267 (2018)
- De la Concepcion, J. C. et al. Polymorphic residues in rice NLRs expand binding and response to effectors of the blast pathogen. Nat. Plants 4, 576-585 (2018).
- Pfeilmeier, S. et al. Expression of the Arabidopsis thaliana immune receptor EFR in Medicago truncatula reduces infection by a root pathogenic bacterium, but not nitrogenfixing rhizobial symbiosis. Plant Biotechnol. J. 17, 569-579 (2019).
- Koller, T. et al. Field grown transgenic Pm3e wheat lines show powdery mildew resistance and no fitness costs associated with high transgene expression. Transgenic Res. 28, 9-20 (2019).
- Xu, G. et al. uORF-mediated translation allows engineered plant disease resistance without fitness costs. Nature 545, 491-494 (2017).
  - The engineering of a pathogen-responsive upstream open reading frame cassette fused to a central Arabidopsis disease-resistance component (NPR1) in transgenic rice creates a crop line that exhibits broad-spectrum disease resistance without a yield penalty.
- Wu, C.-H., Derevnina, L. & Kamoun, S. Receptor networks underpin plant immunity. Science 360, 1300-1301 (2018).
- Hörger, A. C. et al. Balancing selection at the tomato RCR3 Guardee gene family maintains variation in strength of pathogen defense. PLoS Genet. 8, e1002813 (2012)
- Chen, C., E. Z. & Lin, H.-X. Evolution and molecular control of hybrid incompatibility in plants. Front. Plant Sci. 7, 1208 (2016).
- Dangl. J. L., Horvath, D. M. & Staskawicz, B. J. Pivoting the plant immune system from dissection to deployment, Science 341, 746-751 (2013).
- 30 Witek, K. et al. Accelerated cloning of a potato late blight-resistance gene using RenSeq and SMRT sequencing. Nat. Biotechnol. 34, 656-660 (2016).
- Bevan, M. W. et al. Genomic innovation for crop improvement. Nature 543, 346-354 (2017)
- 32. Bailey, P. C. et al. Dominant integration locus drives continuous diversification of plant immune receptors with exogenous domain fusions. Genome Biol. 19, 23 (2018).
- Arora, S. et al. Resistance gene cloning from a wild crop relative by sequence capture and association genetics. Nat. Biotechnol. 37, 139-143 (2019).
- van Schie, C. C. N. & Takken, F. L. W. Susceptibility genes 101: how to be a good host. Annu. Rev. Phytopathol. 52, 551-581 (2014).
- Bezrutczyk, M. et al. Sugar flux and signaling in plant-microbe interactions. Plant J. 93,
  - The engineering of recently identified plant sugar transporters, which provide pathogens with carbohydrates, enables the restriction of disease through reducing the nutrients available to these pathogens.
  - Gomez, M. A. et al. Simultaneous CRISPR/Cas9-mediated editing of cassava eIF4E isoforms nCBP-1 and nCBP-2 reduces cassava brown streak disease symptom severity and incidence. Plant Biotechnol. J. 17, 421-434 (2019).

- Wang, Y. et al. Simultaneous editing of three homoeoalleles in hexaploid bread wheat confers heritable resistance to powdery mildew. Nat. Biotechnol. 32, 947-951 (2014)
- 38. Acevedo-Garcia, J. et al. *mlo*-based powdery mildew resistance in hexaploid bread wheat generated by a non-transgenic TILLING approach. Plant Biotechnol. J. 15,
- Mehta, D. et al. Linking CRISPR-Cas9 interference in cassava to the evolution of editingresistant geminiviruses. Genome Biol. 20, 80 (2019).
- Bebber, D. P., Ramotowski, M. A. T. & Gurr, S. J. Crop pests and pathogens move polewards in a warming world. Nat. Clim. Chang. 3, 985-988 (2013).
- Almeida, R. P. P. et al. Addressing the new global threat of Xylella fastidiosa. Phytopathology **109**, 172-174 (2019).
- 42. Cheng, C. et al. Plant immune response to pathogens differs with changing temperatures, Nat. Commun. 4, 2530 (2013).
- Velásquez, A. C., Castroverde, C. D. M. & He, S. Y. Plant-pathogen warfare under changing 43. climate conditions. Curr. Biol. 28, R619-R634 (2018).
- 44. Chen. S., Zhang, W., Bolus, S., Rouse, M. N. & Dubcovsky, J. Identification and characterization of wheat stem rust resistance gene Sr21 effective against the Ug99 race group at high temperature, PLoS Genet, 14, e1007287 (2018).
  - Resistance (NLR) genes that function at high temperature are effective against a fungal strain (ug99) that causes wheat stem rust disease and is threatening global wheat production.
- Zhang, W. et al. Identification and characterization of Sr13, a tetraploid wheat gene that confers resistance to the Ug99 stem rust race group. Proc. Natl Acad. Sci. USA 114, F9483-F9492 (2017)
- Mickelbart, M. V., Hasegawa, P. M. & Bailey-Serres, J. Genetic mechanisms of abiotic 46. stress tolerance that translate to crop yield stability. Nat. Rev. Genet. 16, 237-251 (2015)
- 47. Ismail, A. M. & Horie, T. Genomics, physiology, and molecular breeding approaches for improving salt tolerance. Annu. Rev. Plant Biol. 68, 405-434 (2017).
- Voesenek, L. A. C. J. & Bailey-Serres, J. Flood adaptive traits and processes: an overview. New Phytol. 206, 57-73 (2015)
- Singh, S., Mackill, D. J. & Ismail, A. M. Physiological basis of tolerance to complete submergence in rice involves genetic factors in addition to the SUB1 gene. AoB Plants 6,
- Xu, K. et al. Sub1A is an ethylene-response-factor-like gene that confers submergence tolerance to rice. Nature 442, 705-708 (2006).
- Fukao, T., Xu, K., Ronald, P. C. & Bailey-Serres, J. A variable cluster of ethylene response 51 factor-like genes regulates metabolic and developmental acclimation responses to submergence in rice. Plant Cell 18, 2021-2034 (2006).
- 52. Dar, M. H. et al. No yield penalty under favorable conditions paving the way for successful adoption of flood tolerant rice, Sci. Rep. 8, 9245 (2018).
- 53. Hattori, Y. et al. The ethylene response factors SNORKEL1 and SNORKEL2 allow rice to adapt to deep water Nature 460, 1026-1030 (2009).
- 54. Kuroha, T. et al. Ethylene-gibberellin signaling underlies adaptation of rice to periodic flooding, Science 361, 181-186 (2018).
  - A loss-of-function allele of SEMIDWARF 1 confers short stature in Green Revolution rice, whereas a functional ethylene-induced allele enables elongation of submerged internodes for grain production above floodwaters
- Ashikari, M. et al. Loss-of-function of a rice gibberellin biosynthetic gene, GA20 oxidase (GA20ox-2), led to the rice 'green revolution'. Breed. Sci. 52, 143-150 (2002).
- Stein, J. C. et al. Genomes of 13 domesticated and wild rice relatives highlight genetic 56. conservation, turnover and innovation across the genus Oryza. Nat. Genet. 50, 285-296 (2018).
  - This inventory of genetic variation in the Oryza genus, and other recent pan-genome studies of crops, empower the exploration of phylogenetic relationships and genetic variation at disease-resistance and stress-resilience loci
- Gibbs, D. J. et al. Homeostatic response to hypoxia is regulated by the N-end rule pathway in plants, Nature 479, 415-418 (2011).
- Licausi, F. et al. Oxygen sensing in plants is mediated by an N-end rule pathway for protein destabilization. Nature 479, 419-422 (2011).
- Mendiondo, G. M. et al. Enhanced waterlogging tolerance in barley by manipulation of expression of the N-end rule pathway E3 ligase PROTEOLYSIS6. Plant Biotechnol. J. 14, 40-50 (2016).
- 60. Lin, C.-C. et al. Regulatory cascade involving transcriptional and N-end rule pathways in rice under submergence, Proc. Natl Acad. Sci. USA 116, 3300-3309 (2019).
  - The submergence tolerance regulator SUB1A is unusual among ethylene-responsive transcription factors-including two related transcription factors encoded by genes it activates—in its insensitivity to oxygen-dependent turnover.
- Liljegren, S. J. et al. SHATTERPROOF MADS-box genes control seed dispersal in Arabidopsis. Nature 404, 766-770 (2000).
- 62 Park, S.-Y. et al. Abscisic acid inhibits type 2C protein phosphatases via the PYR/PYL family of START proteins. Science 324, 1068-1071 (2009).
- Ma, Y. et al. Regulators of PP2C phosphatase activity function as abscisic acid sensors. Science 324, 1064-1068 (2009).
- Zhao, Y. et al. ABA receptor PYL9 promotes drought resistance and leaf senescence. 64. Proc. Natl Acad. Sci. USA 113, 1949-1954 (2016).
- Yang, Z. et al. Leveraging abscisic acid receptors for efficient water use in Arabidopsis. Proc. Natl Acad. Sci. USA 113, 6791-6796 (2016).
- 66. Park, S.-Y. et al. Agrochemical control of plant water use using engineered abscisic acid receptors. Nature 520, 545-548 (2015).
- Mega, R. et al. Tuning water-use efficiency and drought tolerance in wheat using abscisic acid receptors. Nat. Plants 5, 153-159 (2019).
- Kudo, M. et al. A gene-stacking approach to overcome the trade-off between drought stress tolerance and growth in Arabidopsis. Plant J. 97, 240-256 (2019).
- Hu, H. & Xiong, L. Genetic engineering and breeding of drought-resistant crops. Annu. Rev. Plant Biol. 65, 715-741 (2014).

- Décima Oneto, C. et al. Water deficit stress tolerance in maize conferred by expression of an isopentenyltransferase (IPT) gene driven by a stress- and maturation-induced promoter. J. Biotechnol. 220, 66-77 (2016).
- Wang, Y. et al. Molecular tailoring of farnesylation for plant drought tolerance and yield 71. protection. Plant J. 43, 413-424 (2005).
- Saradadevi, R., Palta, J. A. & Siddique, K. H. M. ABA-mediated stomatal response in regulating water use during the development of terminal drought in wheat. Front. Plant Sci. 8, 1251 (2017)
- Uga, Y. et al. Control of root system architecture by DEEPER ROOTING 1 increases rice yield under drought conditions. Nat. Genet. 45, 1097-1102 (2013).
- Arai-Sanoh, Y. et al. Deep rooting conferred by DEEPER ROOTING 1 enhances rice yield in paddy fields. Sci. Rep. 4, 5563 (2014).
- 75. Li, X. et al. Genetic control of the root system in rice under normal and drought stress conditions by genome-wide association study. PLoS Genet. 13, e1006889 (2017).
- Bray, A. L. & Topp, C. N. The quantitative genetic control of root architecture in maize. 76. Plant Cell Physiol. 59, 1919-1930 (2018).
- 77 Bao, Y. et al. Plant roots use a patterning mechanism to position lateral root branches toward available water, Proc. Natl Acad. Sci. USA 111, 9319-9324 (2014).
- 78. Orosa-Puente, B. et al. Root branching toward water involves posttranslational modification of transcription factor ARF7, Science 362, 1407-1410 (2018).
- 79 Hall, A. J. & Richards, R. A. Prognosis for genetic improvement of yield potential and water-limited yield of major grain crops. Field Crops Res. 143, 18-33 (2013).
- Nuccio, M. L., Paul, M., Bate, N. J., Cohn, J. & Cutler, S. R. Where are the drought tolerant crops? An assessment of more than two decades of plant biotechnology effort in crop improvement, Plant Sci. 273, 110-119 (2018).
- Rubio, F., Gassmann, W. & Schroeder, J. I. Sodium-driven potassium uptake by the plant potassium transporter HKT1 and mutations conferring salt tolerance. Science 270, 1660-1663 (1995).
- 82. Mäser, P. et al. Altered shoot/root Na<sup>+</sup> distribution and bifurcating salt sensitivity in Arabidopsis by genetic disruption of the Na<sup>+</sup> transporter AtHKT1. FEBS Lett. 531, 157-161
- Ren, Z.-H. et al. A rice quantitative trait locus for salt tolerance encodes a sodium transporter, Nat. Genet. 37, 1141-1146 (2005).
- Sunarpi, et al. Enhanced salt tolerance mediated by AtHKT1 transporter-induced Na<sup>+</sup> unloading from xylem vessels to xylem parenchyma cells. Plant J. 44, 928-938
- Munns, R. et al. Wheat grain yield on saline soils is improved by an ancestral Na<sup>\*</sup> transporter gene. Nat. Biotechnol. 30, 360-364 (2012).
  - An HKT1 allele that confers sodium tolerance is crossed into a commercial durum wheat variety, thus enhancing yield under salinity stress.
- 86. Barberon, M. et al. Adaptation of root function by nutrient-induced plasticity of endodermal differentiation, Cell 164, 447-459 (2016).
- Duan, L. et al. Endodermal ABA signaling promotes lateral root quiescence during salt stress in Arabidonsis seedlings. Plant Cell 25, 324-341 (2013).
  - Root growth dynamics under salinity stress are regulated by processes in a specific cell layer, and effective gene engineering of these dynamics may require use of promoters with defined spatial and temporal activity.
- Huang, C. F. et al. A bacterial-type ABC transporter is involved in aluminum tolerance in rice. Plant Cell 21, 655-667 (2009).
- 89 Maron, L. G. et al. Aluminum tolerance in maize is associated with higher MATE1 gene copy number. Proc. Natl Acad. Sci. USA 110, 5241-5246 (2013).
- Pallotta, M. et al. Molecular basis of adaptation to high soil boron in wheat landraces and elite cultivars. Nature 514, 88-91 (2014).
  - The molecular characterization of loci associated with boron tolerance uncovers variation in gene copy number and expression patterns that translate to the field.
- Horton, R. M., Mankin, J. S., Lesk, C., Coffel, E. & Raymond, C. A review of recent advances in research on extreme heat events. Curr. Clim. Change Rep. 2, 242-259 (2016).
- Moore, F. C., Baldos, U., Hertel, T. & Diaz, D. New science of climate change impacts on agriculture implies higher social cost of carbon. Nat. Commun. 8, 1607 (2017).
- Lobell, D. B. et al. Greater sensitivity to drought accompanies maize yield increase in the U.S. Midwest. Science 344, 516-519 (2014).
- Zhao, C. et al. Temperature increase reduces global yields of major crops in four independent estimates, Proc. Natl Acad. Sci. USA 114, 9326-9331 (2017).
- Bita, C. E. & Gerats, T. Plant tolerance to high temperature in a changing environment: scientific fundamentals and production of heat stress-tolerant crops, Front, Plant Sci. 4. 273 (2013).
- Jackson, R. D., Idso, S. B., Reginato, R. J. & Pinter, P. J. Jr. Canopy temperature as a crop 96. water stress indicator, Wat, Resour, Res. 17, 1133-1138 (1981).
- Cooper, M., Gho, C., Leafgren, R., Tang, T. & Messina, C. Breeding drought-tolerant maize hybrids for the US corn-belt: discovery to product. J. Exp. Bot. 65, 6191-6204 (2014).
- 98. Nuccio, M. L. et al. Expression of trehalose-6-phosphate phosphatase in maize ears improves yield in well-watered and drought conditions. Nat. Biotechnol. 33, 862-869 (2015)
  - Improving drought resilience of maize yield was achieved through the design of a promoter that targets maize-ear tissue in a stage-specific manner
- Kretzschmar, T. et al. A trehalose-6-phosphate phosphatase enhances anaerobic germination tolerance in rice. Nat. Plants 1, 15124 (2015)
- 100. Griffiths, C. A. et al. Chemical intervention in plant sugar signalling increases yield and resilience. Nature 540, 574-578 (2016).
- The effective delivery of membrane-permeable and activatable small molecules could provide a new management strategy for drought protection. 101. Zhu, X.-G., Long, S. P. & Ort, D. R. Improving photosynthetic efficiency for greater yield.
- Annu. Rev. Plant Biol. 61, 235-261 (2010). 102. Dann, M. & Leister, D. Enhancing (crop) plant photosynthesis by introducing novel genetic
- diversity. Phil. Trans. R. Soc. Lond. B 372, 20160380 (2017).
- 103. Blankenship, R. E. Early evolution of photosynthesis. Plant Physiol. 154, 434-438 (2010)

# Review

- 104. Farquhar, G. D., von Caemmerer, S. & Berry, J. A. A biochemical model of photosynthetic CO<sub>2</sub> assimilation in leaves of C<sub>3</sub> species. *Planta* 149, 78-90 (1980).
- Long, S. P., Marshall-Colon, A. & Zhu, X.-G. Meeting the global food demand of the future by engineering crop photosynthesis and yield potential. Cell 161, 56–66 (2015).
- Ort, D. R. et al. Redesigning photosynthesis to sustainably meet global food and bioenergy demand. Proc. Natl Acad. Sci. USA 112, 8529–8536 (2015).
- 107. Evans, J. R. & Clarke, V. C. The nitrogen cost of photosynthesis. J. Exp. Bot. 70, 7-15 (2019).
- 108. Murchie, E. H. & Niyogi, K. K. Manipulation of photoprotection to improve plant photosynthesis. *Plant Physiol.* **155**, 86–92 (2011).
- Betti, M. et al. Manipulating photorespiration to increase plant productivity: recent advances and perspectives for crop improvement. J. Exp. Bot. 67, 2977–2988 (2016).
- Simkin, A. J., McAusland, L., Lawson, T. & Raines, C. A. Overexpression of the RieskeFeS protein increases electron transport rates and biomass yield. *Plant Physiol.* 175, 134–145 (2017).
- von Caemmerer, S., Quick, W. P. & Furbank, R. T. The development of C<sub>4</sub> rice: current progress and future challenges. Science 336, 1671–1672 (2012).
- Evans, L. T. Adapting and improving crops: the endless task. Phil. Trans. R. Soc. Lond. B 352, 901–906 (1997).
- Sharwood, R. E. Engineering chloroplasts to improve Rubisco catalysis: prospects for translating improvements into food and fiber crops. New Phytol. 213, 494–510 (2017).
- Salesse-Smith, C. E. et al. Overexpression of Rubisco subunits with RAF1 increases Rubisco content in maize. Nat. Plants 4, 802–810 (2018).
- Driever, S. M. et al. Increased SBPase activity improves photosynthesis and grain yield in wheat grown in greenhouse conditions. Phil. Trans. R. Soc. Lond. B 372, 20160384 (2017).
- 116. Kromdijk, J. et al. Improving photosynthesis and crop productivity by accelerating recovery from photoprotection. Science 354, 857–861 (2016).
  Transgenic tobacco with accelerated interconversion of violaxanthin and zeaxanthin in the xanthophyll cycle, and increased amounts of a photosystem II subunit, yielded 15% greater production of plant biomass in natural field conditions.
- 117. South, P. F., Cavanagh, A. P., Liu, H. W. & Ort, D. R. Synthetic glycolate metabolism pathways stimulate crop growth and productivity in the field. Science 363, eaat9077 (2019).
  Transgenic tobacco with synthetic glycolate metabolism in the chloroplast showed enhanced photosynthesis and increased biomass in the field.
- Zhang, J. et al. Insights into the molecular mechanisms of CO<sub>2</sub>-mediated regulation of stomatal movements. Curr. Biol. 28, R1356–R1363 (2018).
- Hsu, P.-K. et al. Abscisic acid-independent stomatal CO<sub>2</sub> signal transduction pathway and convergence of CO<sub>2</sub> and ABA signaling downstream of OST1 kinase. Proc. Natl Acad. Sci. USA 115, E9971–E9980 (2018).
- Hu, H. et al. Carbonic anhydrases are upstream regulators of CO<sub>2</sub>-controlled stomatal movements in guard cells. Nat. Cell Biol. 12, 87–93 (2010).
  - A strategy for improving water use efficiency via modulating the stomatal CO<sub>2</sub> response is shown.
- Franks, P. J. et al. Sensitivity of plants to changing atmospheric CO<sub>2</sub> concentration: from the geological past to the next century. New Phytol. 197, 1077–1094 (2013).
- Hughes, J. et al. Reducing stomatal density in barley improves drought tolerance without impacting on yield. *Plant Physiol.* 174, 776–787 (2017).
- 123. Caine, R. S. et al. Rice with reduced stomatal density conserves water and has improved drought tolerance under future climate conditions. New Phytol. 221, 371–384 (2019).
- 124. Głowacka, K. et al. Photosystem II subunit S overexpression increases the efficiency of water use in a field-grown crop. Nat. Commun. 9, 868 (2018).
- Rogers, A., Ainsworth, E. A. & Leakey, A. D. B. Will elevated carbon dioxide concentration amplify the benefits of nitrogen fixation in legumes? *Plant Physiol.* 151, 1009–1016 (2009).
- Schroeder, J. I. et al. Using membrane transporters to improve crops for sustainable food production. Nature 497, 60–66 (2013).
- Li, S. et al. Modulating plant growth-metabolism coordination for sustainable agriculture. Nature 560, 595–600 (2018).
- 128. Cormier, F. et al. Breeding for increased nitrogen-use efficiency: a review for wheat (T. aestivum L.). Plant Breed. 135, 255–278 (2016).
- Oldroyd, G. E. D. Speak, friend, and enter: signalling systems that promote beneficial symbiotic associations in plants. Nat. Rev. Microbiol. 11, 252–263 (2013).
- Harrison, M. J. Signaling in the arbuscular mycorrhizal symbiosis. Annu. Rev. Microbiol. 59, 19–42 (2005).
- Choi, J., Summers, W. & Paszkowski, U. Mechanisms underlying establishment of arbuscular mycorrhizal symbioses. Annu. Rev. Phytopathol. 56, 135–160 (2018).
- Foo, E., Yoneyama, K., Hugill, C. J., Quittenden, L. J. & Reid, J. B. Strigolactones and the regulation of pea symbioses in response to nitrate and phosphate deficiency. *Mol. Plant* 6, 76–87 (2013).
- Yang, J. et al. Polyprotein strategy for stoichiometric assembly of nitrogen fixation components for synthetic biology. Proc. Natl Acad. Sci. USA 115, E8509–E8517 (2018).
- Burén, S. & Rubio, L. M. State of the art in eukaryotic nitrogenase engineering. FEMS Microbiol. Lett. 365, fnx274 (2018).
- Griesmann, M. et al. Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis. Science 361, eaat1743 (2018).
- van Velzen, R. et al. Comparative genomics of the nonlegume Parasponia reveals insights into evolution of nitrogen-fixing rhizobium symbioses. Proc. Natl Acad. Sci. USA 115, E4700–E4709 (2018).
- Busby, P. E. et al. Research priorities for harnessing plant microbiomes in sustainable agriculture. PLoS Biol. 15, e2001793 (2017).

- Edwards, J. A. et al. Compositional shifts in root-associated bacterial and archaeal microbiota track the plant life cycle in field-grown rice. PLoS Biol. 16, e2003862 (2018).
- 139. Castrillo, G. et al. Root microbiota drive direct integration of phosphate stress and immunity. *Nature* **543**, 513–518 (2017).
- Pieterse, C. M. J. et al. Induced systemic resistance by beneficial microbes. Annu. Rev. Phytopathol. 52, 347–375 (2014).
- Kwak, M.-J. et al. Rhizosphere microbiome structure alters to enable wilt resistance in tomato. Nat. Biotechnol. 36, 1100–1109 (2018).
   Comparative rhizosphere metagenomics enables the identification of protective
  - Comparative rhizosphere metagenomics enables the identification of protective microbial strains from the root-associated microbiota of disease-resistant tomato plants.
- Subramanian, K. S., Charest, C., Dwyer, L. M. & Hamilton, R. I. Arbuscular mycorrhizas and water relations in maize under drought stress at tasselling. *New Phytol.* 129, 643–650 (1995).
- Augé, R. M. Water relations, drought and vesicular-arbuscular mycorrhizal symbiosis. Mycorrhiza 11. 3–42 (2001).
- 144. Fitzpatrick, C. R. et al. Assembly and ecological function of the root microbiome across angiosperm plant species. *Proc. Natl Acad. Sci. USA* **115**, E1157–E1165 (2018).
- Hiruma, K. et al. Root endophyte Colletotrichum tofieldiae confers plant fitness benefits that are phosphate status dependent. Cell 165, 464–474 (2016).
- 146. Van Deynze, A. et al. Nitrogen fixation in a landrace of maize is supported by a mucilage-associated diazotrophic microbiota. PLoS Biol. 16, e2006352 (2018).
  In this study, a maize landrace is demonstrated to stimulate symbiosis with microbes
  - that have nitrogenase activity, which contributes to nitrogen nutrition.
    7. Challinor, A. J., Koehler, A.-K., Ramirez-Villegas, J., Whitfield, S. & Das, B. Current warming
- will reduce yields unless maize breeding and seed systems adapt immediately. *Nat. Clim. Chang.* **6**, 954–958 (2016).
- Atlin, G. N., Cairns, J. E. & Das, B. Rapid breeding and varietal replacement are critical to adaptation of cropping systems in the developing world to climate change. Glob. Food Sec. 12, 31–37 (2017).
- Walsh, M. J. et al. Algal food and fuel coproduction can mitigate greenhouse gas emissions while improving land and water-use efficiency. *Environ. Res. Lett.* 11, 114006 (2016).
- Dempewolf, H. et al. Adapting agriculture to climate change: a global initiative to collect, conserve, and use crop wild relatives. Agroecol. Sustain. Food Syst. 38, 369–377 (2014).
- 151. McCouch, S. et al. Agriculture: feeding the future. Nature 499, 23-24 (2013).
- Khanday, I., Skinner, D., Yang, B., Mercier, R. & Sundaresan, V. A male-expressed rice embryogenic trigger redirected for asexual propagation through seeds. *Nature* 565, 91–95 (2019).
- Becker, R. A. & Wilks, A. R. Maps: draw geographical maps, v.3.3.0, https://cran.r-project. org/package=maps/(2018).
- Brakenridge, G. R. Global Active Archive of Large Flood Events (Dartmouth Flood Observatory, University of Colorado), http://floodobservatory.colorado.edu/Archives/ index.html (accessed 2019).
- López-Calcagno, P. E. et al. Overexpressing the H-protein of the glycine cleavage system increases biomass yield in glasshouse and field-grown transgenic tobacco plants. Plant Biotechnol. J. 17, 141–151 (2019).
- 156. Lin, M. T., Occhialini, A., Andralojc, P. J., Parry, M. A. J. & Hanson, M. R. A faster Rubisco with potential to increase photosynthesis in crops. *Nature* 513, 547–550 (2014).

Acknowledgements We apologize to those authors whose research could not be cited owing to space limits. We thank P. J. Franks and E. Buckler for discussions, and A. Digrado and C. Benjamin for assistance with figures. Research in the authors' laboratories was supported by grants from the US National Science Foundation to J.I.S. (MCB-1900567) and J.B.-S. (IOS-1546879; IOS-1810468; IOS-1856749); the US National Institutes of Health to J.I.S. (GM060396-ES010337); the National Institute of Food and Agriculture to J.B.-S. (2017-67013-26194; 2019-67013-29313); the Max-Planck Society and the Deutsche Forschungsgemeinschaft (DFG; German Research Foundation) under Germany's Excellence Strategy (EXC-2048/1, project 390686111 and Priority Program 2125 'DECRyPT') to J.E.P.; a sub-award to E.A.A. from the University of Illinois as part of the Realizing Increased Photosynthetic Efficiency (RIPE) project (OPP1060461), UK AlD and the Foundation for Food and Agricultural Research; and the Engineering the Nitrogen Symbiosis in Africa project (OPP172165) sponsored by the Bill & Melinda Gates Foundation and Gatsby Foundation grants to G.E.D.O.

 $\textbf{Author contributions} \ \textbf{All authors contributed equally to this work.}$ 

Competing interests The authors declare no competing interests.

#### Additional information

Correspondence and requests for materials should be addressed to J.B.-S. or J.I.S.

Peer review information Nature thanks Martin van Ittersum, Pamela Ronald and Cyril Zipfel for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints. Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© Springer Nature Limited 2019

# Immunization: vital progress, unfinished agenda

https://doi.org/10.1038/s41586-019-1656-7

Received: 5 June 2019

Accepted: 6 September 2019

Published online: 6 November 2019

Peter Piot1\*, Heidi J. Larson12,3, Katherine L. O'Brien4, John N'kengasong5, Edmond Ng1, Samba Sow<sup>6</sup> & Beate Kampmann<sup>1,7</sup>

Vaccination against infectious diseases has changed the future of the human species. saving millions of lives every year, both children and adults, and providing major benefits to society as a whole. Here we show, however, that national and sub-national coverage of vaccination varies greatly and major unmet needs persist. Although scientific progress opens exciting perspectives in terms of new vaccines, the pathway from discovery to sustainable implementation can be long and difficult, from the financing, development and licensing to programme implementation and public acceptance. Immunization is one of the best investments in health and should remain a priority for research, industry, public health and society.



On 14 May 1796, 73 years before the first issue of *Nature*, and inspired by Lady Montagu's "variolation" concept, Edward Jenner inoculated eight-year-old James Phipps with cowpox pus to prove that the less virulent cowpox would protect against smallpox. This experiment was a game changer in medicine and health. For the first time, it was possible to medically prevent infection in a healthy person. Although vaccines have been widely introduced in high-income countries since the late 1950s, it took 180 years after Jenner before the Expanded Programme on Immunization (EPI) was launched in 1974, promoting access to six essential vaccines in all countries worldwide. Today, vaccines against 26 infectious diseases are internationally available according to the World Health Organization (WHO)<sup>1</sup>, although more have been licensed worldwide, changing the future of the human species. Others are in experimental public health use, such as Ebola vaccines, or pilot implementation such as the RTS, S malaria vaccine, and about 240 vaccine candidates are in development<sup>2</sup> (Table 1). The US Centers for Disease Control and Prevention declared vaccination the number one success story for public health in the twentieth century<sup>3</sup>.

However, progress in vaccine coverage remains highly uneven-both between and within countries—which threatens hard-won progress and raises uncertainty about how to make further advances. Vaccinepreventable diseases such as measles are on the rise, and episodes of vaccine reluctance and refusal are occurring globally, questioning one of the most transformative interventions for survival and health.

This Review focuses on preventive immunization in humans and its impact (rather than on the vaccines themselves), including in low-, middle- and high-income countries. We discuss the current status of vaccine coverage, as well as unmet needs, four hurdles to overcome to ensure sustainable immunization programmes starting with the discovery of a new vaccine, the growing issue of vaccine confidence, and conclude with several opportunities and needed actions to ensure the full potential of immunization for human health and society. Developmental challenges for vaccine production for low- and middle-income countries, which were recently discussed in separate articles<sup>4,5</sup>, and therapeutic vaccines are not discussed.

Vaccines are biological products that induce protective immunity against infection and disease; they consist of sub-components, killed or inactivated organisms or live-attenuated viruses that train the immune system for a future response to a natural infection. They are probably the only medical intervention that is recommended for every single individual on the planet. Unlike therapeutics, vaccines are used in healthy people, and demand a very high standard of safety and require continuous monitoring for potential side effects. Besides considerations of safety, effectiveness, impact and cost, this raises complex governance, regulatory and public trust issues. All countries have a national immunization plan, often with goals inspired by the Global Vaccine Action Plan (GVAP) framework for 2011–20206.

#### How immunization has crucially benefited society

It is hard to imagine a world without vaccines. A decade ago, the WHO, UNICEF and the World Bank estimated that routine childhood immunization programmes were preventing more than 2.5 million deaths every year<sup>7</sup>. With the increase in vaccine coverage, the growth of populations, and the introduction of new life-saving vaccines, immunization is ever more important for survival. In addition to preventing deaths, vaccines prevent disease and disability, including in adults and the elderly. In a high-income country such as the United States, for a single birth cohort, vaccines prevent nearly 20 million cases of disease, and more than 40,000 deaths8.

A vaccine has for the first time in history eradicated a human disease, smallpox. Efforts to eradicate polio are in the final stages, with

Office of the Director, Vaccine Centre and Vaccine Confidence Project, London School of Hygiene & Tropical Medicine, London, UK. 2Institute for Health Metrics and Evaluation, University of Washington Seattle WA USA 3Centre for the Evaluation of Vaccination (CEV) University of Antwern Relgium 4Department of Immunization Vaccines and Riologicals World Health Organization, Geneva, Switzerland. 5 Africa Centres for Disease Control and Prevention, Addis Ababa, Ethiopia. 6 Center for Vaccine Development, Bamako, Mali. 7 MRC Unit The Gambia at the LSHTM, Banjul, The Gambia, \*e-mail: peter.piot@lshtm.ac.uk

#### Review

#### Table 1 | Historic timeline of introduction of vaccines

Year	Disease	Year	Disease
1798	Smallpox	1992	Japanese encephalitis (mouse brain)
1885	Rabies	1993	Cholera (recombinant toxin B)
1896	Cholera	1994	Typhoid (Vi) polysaccharide
1896	Typhoid	1994	Cholera (attenuated)
1897	Plague	1995	Varicella
1923	Diphtheria toxoid	1996	Hepatitis A
1926	Pertussis (WC)	1996	Pertussis (acellular)
1926	Tetanus toxoid	1998	Lyme OspA
1927	Tuberculosis (BCG)	1999	Meningococcal conjugate (group C) <sup>a</sup>
1935	Yellow fever	1999	Rotavirus (reassortant)
1936	Influenza	2000	Pneumococcal conjugate (7-valent) <sup>a</sup>
1937	Tickborne encephalitis	2003	Influenza (intranasal, cold-adapted)
1938	Typhus	2005	Meningococcal conjugates (4-valent) <sup>a</sup>
1955	Polio (inactivated)	2006	Human papillomavirus recombinant (4-valent)
1963	Measles	2006	Rotavirus (attenuated and new reassortants)
1963	Polio (oral)	2006	Varicella Zoster
1967	Mumps	2008	Rotavirus (monovalent)
1969	Rubella	2009	Japanese encephalitis (Vero cell)
1970	Anthrax secreted proteins	2009	Cholera (WC only)
1974	Meningococcus polysaccharide	2009	Human papillomavirus recombinant (2-valent)
1977	Pneumococcus polysaccharide (14-valent)	2010	Meningococcal type A conjugate (monovalent)
1980	Adenovirus	2010	Pneumococcal conjugate (13-valent)
1980	Rabies (cell culture)	2014	Human papillomavirus (9-valent)
1981	Tickborne encephalitis	2014	Meningococcal type B (fH factor)
1981	Hepatitis B (plasma derived)	2015	Ebola (unlicensed) <sup>b</sup>
1983	Pneumococcus polysaccharide (23-valent)	2015	Malaria <sup>c</sup>
1985	Haemophilus influenzae type b polysaccharide	2015	Dengue
1986	Hepatitis B surface antigen recombinant	2015	Meningococcal type B <sup>d</sup>
1987	Haemophilus influenzae type b conjugate <sup>a</sup>	2016	Cholera (oral)
1989	Typhoid (Salmonella Ty21a)	2018	Typhoid conjugate <sup>a</sup>
1991	Cholera (WC-rBS)		

Table adapted from Plotkin & Plotkin (2018)<sup>111</sup>. The year of licensing is indicated wherever possible. rBS, recombinant B subunit; WC, whole cell.

only two countries, Afghanistan and Pakistan, still experiencing wild transmission of the polio virus. All countries with the exception of 13 have eliminated neonatal and maternal tetanus. Without vaccination, there would be far more infections that require antibiotic therapy, exacerbating the major problem of drug-resistant infections.

Between 1990 and 2017, immunization contributed to a 55% global decline in under-five mortality rates, with a drop from 87 to 39 deaths per 1,000 live births $^9$ . More than 14 million deaths are estimated to have been prevented by measles vaccination alone between 2011 and 2020 $^6$ 

Vaccination benefits not only those who are vaccinated, but also others in their family and community. This population-wide benefit, known as 'herd immunity', reduces the exposure of unvaccinated individuals to pathogens through a reduction or interruption of the chains of transmission. A recent study in Kenya showed that the introduction of a pneumococcal vaccine resulted in not only a major reduction in invasive pneumococcal disease, but also a nearly 100% decline in incidence among infants too young to be vaccinated, and a more than 74% reduction among unvaccinated children 10. Community or herd immunity is an

important consideration when estimating the full public health value of immunization. The threshold to achieve such community protection can be as high as 95% for measles, but as low as 80% for rubella, and 60% in high-income settings for the effect to begin for pneumococcal vaccination, which means that the programme strength required to derive additional impact varies substantially by vaccine  $^{\rm 11-13}$ . These differences in the required critical vaccination coverage rates are due to the basic reproductive ratio of an infection  $(R_0)^{\rm 14}$ , which can vary greatly among various infectious diseases. The  $R_0$  of a specific infection indicates the average number of cases one case generates in a population—in the case of measles it is 12–18, which is among the highest  $^{\rm 15}$ . It is an indicator of how contagious an infection is, and determines the minimum level of vaccination coverage needed to generate herd immunity.

Potential long-term effects beyond direct protection against a specific pathogen or disease have been attributed to several vaccines, in particular the BCG vaccine against tuberculosis and the measles vaccine, in which observational studies suggested a survival advantage compared with children who had remained unvaccinated. These non-specific effects (also known as heterologous effects) would add to the disease-specific, proven

<sup>&</sup>lt;sup>a</sup>Capsular polysaccharide conjugated to carrier proteins.

<sup>&</sup>lt;sup>b</sup>An investigational vaccine, rVSV-ZEBOV, was used under 'expanded access' during the Ebola outbreak in West Africa in 2015 and the 2018-2019 outbreak in the Democratic Republic of the Congo; the Ad26.ZEBOV/MVA-BN-Filo vaccine was used in 2019 in Rwanda and the Democratic Republic of the Congo.

<sup>°</sup>Positive opinion from the EMA under article 58 issued in 2015. Approved for routine use in pilot implementation settings in Ghana, Malawi and Kenya in 2018

dReverse vaccinology.

Table 2 | Vaccines across the human life cycle

Recommended immunization schedule	Vaccines	
Life cycle stage		
Newborns	BCG; hepatitis B; polio.	
Infants/toddlers	Diphtheria; tetanus; pertussis; polio; <i>Haemophilus influenza</i> type b; hepatitis B; influenza; pneumococcus; rotavirus; malaria; meningococcus; varicella; measles; mumps; rubella; typhoid; yellow fever. Under development: RSV; <i>Salmonella</i> spp.; <i>Shigella</i> spp.; ETEC.	
Older children and adolescents	HPV; influenza; meningococcus; diphtheria booster; tetanus booster; pertussis booster. Under development: group A streptococcus.	
Adults	Influenza; diphtheria booster; tetanus booster; pertussis booster; varicella; HPV (depending on age at initial vaccination).	
Pregnant women	Tetanus; influenza; pertussis. Under development: group B streptococcus; RSV; CMV.	
Older adults (≥65 years)	Influenza; diphtheria booster; tetanus booster; pertussis booster; pneumococcus; shingles.	
Special health conditions (adults)		
Immuno-compromised (including HIV infection) <sup>a</sup>	Influenza; pneumococcus.	
HIV infection <sup>a</sup>	Influenza; pneumococcus; hepatitis B; meningococcus. For CD4 count ≥200 cells per µl: measles; mumps; rubella; varicella.	
Asplenia, complement disorder <sup>b</sup>	olement disorder <sup>b</sup> Influenza; pneumococcus; meningococcus; <i>Haemophilus influenza</i> type b.	
Chronic kidney disease (including haemodialysis <sup>b</sup>	Influenza; pneumococcus; hepatitis B.	
Chronic liver disease <sup>b</sup>	Influenza; pneumococcus; hepatitis A; hepatitis B.	
Diabetes <sup>b</sup>	Influenza; pneumococcus; hepatitis B.	
Heart or lung disease <sup>b</sup>	Influenza; pneumococcus.	
Other circumstances		
Travel	Hepatitis A; hepatitis B; typhoid, rabies; yellow fever; Japanese encephalitis; cholera; meningococcus; malaria.	
Healthcare workers	Hepatitis B; influenza; measles; mumps; rubella; varicella; diphtheria; tetanus; pertussis; polio; BCG.	

Not only do vaccines provide important health benefits for all stages in life, but they also provide benefits for travellers, healthcare workers and individuals with existing health conditions. Note that the lists of vaccines are illustrative only, rather than exhaustive, and do not indicate that these are universally recommended for each life phase in all countries. Routine vaccines recommended by the WHO are available at: https://www.who.int/immunization/policy/immunization tables/en/ (last accessed 3 September 2019), BCG. Bacillus Calmette-Guérin: CMV: cytomegalovirus; ETEC, enterotoxigenic Escherichia coli; RSV, respiratory syncytial virus

benefits of vaccines, and have been attributed to epigenetic changes in innate immune cells as opposed to the adaptive immunity induced by the antigen-specific responses to the vaccine 16,17. However, the importance of heterologous effects remains controversial, and plausible immunological findings still need to be validated in large-scale clinical trials.

The benefits of vaccines in general go beyond health, and include economic, educational, health security and other benefits<sup>18</sup>. Their full economic value is not sufficiently quantified in assessments of costbenefit, or in investment terms, and is an increasing area of inquiry and empiric measurement19.

Vaccination is a sound investment. Thus, the return on investment from childhood immunization in low- and middle-income countries is high. For every US\$1 invested in immunization against ten diseases, \$16-\$18 are saved in healthcare costs, and the net return is as high as \$44 per dollar spent when the broad economic benefits are considered, although the return on the investment varies by individual vaccine<sup>20</sup>. This is compared with the cost per DTPcv3-vaccinated child of \$27 (having received all three doses of diphtheria-tetanus-pertussis (DTP)-containing vaccine)21. In the United States, the net economic benefits of vaccination in one birth cohort amount to almost \$69 million<sup>22</sup>.

Modelling and observational data suggest that in low- and middleincome countries, vaccination contributes to the alleviation of, and protection against, poverty. Financial risk protection provided by the benefits of vaccination are accrued by the poorest households by the reduction of catastrophic and impoverishing health expenditures 23,24. There is also evidence that vaccination improves childhood physical development, educational outcomes, and equity in distribution of health gains<sup>25</sup>. Finally, without vaccines, absenteeism from school and work would be much higher, and periodic epidemics would disrupt society. The economic effects of periodic influenza epidemics, for example, are enormous<sup>26–28</sup>, and can be reduced by immunization<sup>29</sup>.

#### Vaccination is a lifetime investment

In addition to being the backbone of maternal and child health, vaccines provide important health benefits for all stages in life (Table 2). Given adaptations of the immune system throughout life, not all vaccines work equally well at all stages of life or in all geographical regions30,31.

Starting in infancy, the presence of maternal antibodies in the newborn can impede the response to vaccines, as the neonatal immune system undergoes its own journey of ontogeny, which enables it to adapt from the 'sterile' in utero environment to the confrontation with colonizing and potentially pathogenic microorganisms<sup>32</sup>. Particular immunological pathways have been identified<sup>33</sup>.

Despite considerable progress in reducing the rates of under-five mortality, important gaps remain in addressing neonatal morbidity and mortality. Neonates are particularly vulnerable to infection with Gramnegative bacteria and group B streptococcus, for which no neonatal vaccines currently exist<sup>33,34</sup>. The gap in early protection can potentially be bridged by administering vaccines to women in pregnancy, relying on passively transferred antibodies to protect infants in the first few months of life, until vaccinations administered in infancy or later can provide protection. On the basis of this principle, tetanus, influenza and pertussis vaccinations are recommended for pregnant women to prevent neonatal infections such as neonatal tetanus<sup>35</sup>. This maternal

The following vaccines are recommended for these conditions in the United States: diphtheria booster: tetanus booster: pertussis booster.

<sup>&</sup>lt;sup>b</sup>The following vaccines are recommended for these conditions in the United States: measles; mumps; rubella; varicella.

#### Review

#### Table 3 | From discovery to sustainable effect of immunization: overcoming four major hurdles

	Issues	Selected actions needed
First hurdle: from discovery to early clinical development	Few discoveries make it to actual products     High risk for companies     Safety key issue	Incentives for industry for vaccines with no market in high-income countries     Public–private partnerships and philanthropy
Second hurdle: from early clinical development to large efficacy trials	Very expensive—two-thirds of total costs of new vaccine development     Particularly challenging for vaccine candidates without high-income market potential     Safety major issue, besides immunogenicity and efficacy     Complex road to licensing     Can take 3–10 years or longer	End-to-end product planning need for major boost from private and public funding     Clinical trial capacity and rationalizing trial methodology     Regulatory harmonization and speed     Manufacturing availability for GMP products to be used in trials
Third hurdle: from vaccine licensure to broad scale implementation	Dependent on policy recommendations, cost-effectiveness deliberations and political priority  Country capacity to take on new vaccines; that is, human and financial resources and the time to build political support and community demand  Logistical issues—for example, cold chain, procurement management, organization of vaccination to ensure equity of access  Supply not always sufficient  Highly variable timeline by country	End-to-end product solution     National and international funding, Gavi transition management, tendering processes     National regulatory harmonization     Policy clarification and political leadership     Manufacturing capacity     Research on full societal value of vaccine assessment, implementation research and relevant costeffectiveness models     Equity of access
Fourth hurdle: achieving consistent, long-term supply and demand sustainability	Continuing concern for every national immunization programme Issues may arise even after years of implementation Complex interplay of service delivery, supply and demand, societal trust, political and humanitarian conflicts Never ending	Policy and political commitment Sustainable funding Management and logistics Tender processes Manufacturing capacity Good communication, safety surveillance and vigilance including promptly addressing safety signals and signs of vaccine hesitancy

immunization strategy may be expanded with promising vaccines against group B streptococcus and respiratory syncytial virus<sup>36</sup>.

For adolescents, life-saving vaccines against human papilloma virus (HPV; the cause of cervical, anal, penile and head and neck cancers) are being increasingly introduced and need to be administered before the likely acquisition of HPV via sexual contacts. Vaccines against meningococcal meningitis—a potentially lethal infection with a second peak in adolescence—have also been introduced into this age group in some countries. New platforms such as schools had to be engaged to administer these vaccines.

Outbreaks of mumps have very occasionally been seen in teenagers, despite a solid vaccination record. This highlights the need for surveillance of all age groups for disease outbreaks, and could be due to waning of protection induced by vaccines that are otherwise regarded as highly efficacious<sup>37–39</sup>.

Booster vaccines against diphtheria, tetanus and polio are required to guarantee long-lasting protection and are required throughout adulthood to maintain protective immunity levels—although recommendations may vary by country.

A life-course approach to vaccination has become ever more pressing with pneumonia, influenza and shingles differentially affecting older adults, and death rates from pneumonia and influenza 130 times higher for adults over 85 than for younger adults 40. Vaccination of the elderly with existing vaccines could prevent up to 90,000 deaths per year in the United States alone<sup>41</sup>.

Adultimmunizationdoesnothaveaclearprioritizationinlow-andmiddleincome countries, and is a complex programme across high-income countries. It is different from paediatric immunization, which has a global programme and focused, substantial funding. As the demographics are shifting across the world to an older distribution, a focus on adult immunization will become increasingly relevant, as advocated by the World Coalition on Adult Vaccination<sup>42</sup>. Despite national recommendations<sup>43,44</sup>, vaccine coverage among adults in high-income countries is uneven<sup>45</sup> (vaccine coverage for herpes zoster, which causes shingles, among adults aged 60 or over in the United States was 24% compared with 65% for influenza among those aged 65 or over), and very low or not even available in most low- and middle-income countries<sup>46</sup>. Yet, several studies have shown good cost-effectiveness of adult vaccinations against influenza. pneumococcalinfection, shingles, HPV and tetanus-diphtheria-pertussis<sup>47</sup>.

Important gaps also exist in our understanding of the fundamental biology of adult immunization. Owing to 'immunosenescence'-the gradual decline of the immune system associated with ageing-vaccination of older adults is in general not as effective as in younger people, but the reasons for poorer responsiveness are not well defined, and require a new effort in terms of strategies and products for immunization of adults. However, it is likely that several compartments of the immune system are affected<sup>48</sup>.

There are three areas in which alterations to increase vaccine efficacy in the elderly could be considered: (i) increased vaccine potency; (ii) the use of adjuvants to enhance immunity; and (iii) application of immune modulators or other interventions to alter host immunity generally.

As populations age across the world, it will be increasingly important to identify how to integrate immunization programmes in health and care services to reach all age groups.

In addition, vaccinations are needed for travel, particular professions or specific health conditions <sup>49–51</sup>, and international travel has had a role in the resurgence of measles in areas such as the United States<sup>52</sup>.

#### From discovery to impact: four hurdles to overcome

There are still major infectious diseases that required an effective vaccine for control and ultimate elimination, such as HIV infection and tuberculosis. Therefore, the continuing development of new vaccines is a public health imperative. Unfortunately, most early vaccine candidates in the discovery phase never make it as a safe and effective product. Development and deployment of vaccines is a long and complex process. We briefly describe here four hurdles that need to be overcome from the discovery phase of a new vaccine to sustainable population impact (Table 3).

The first hurdle is a 'valley of death' from discovery to early clinical development, when a potential antigen, adjuvant or new vaccine formulation developed in the laboratory is further tested for clinical proofof-concept and safety in humans, in addition to optimizing production elements. Real progress has been made in recent years owing to several

public and private initiatives that are helping partly to overcome this first major challenge, such as the Coalition for Epidemic Preparedness Innovation (CEPI)<sup>53</sup>, which was created after the 2014–2015 Ebola epidemic in West Africa to accelerate the development of vaccines against epidemic pathogens<sup>2,4,54</sup>.

The second hurdle in vaccine development, also referred to as the 'second valley of death', relates to the shift from early clinical development to the large and very expensive efficacy trials most often needed<sup>4</sup>, unless a previous similar vaccine is already developed and a new product can be licensed using an established correlate of immunity or protection. This is also the most expensive phase of vaccine development, absorbing more than two-thirds of the total costs of development of a new vaccine, including the building of special manufacturing facilities and conducting phase 3 trials in several countries. ideally with independent research partners. Often, this major financial effort is beyond the means of smaller biotech companies, and in general only big pharmaceutical companies and large foundations or public institutions have the financial bandwidth to support such trials that can cost as much as hundreds of millions of dollars. For vaccine candidates without a prospect of a high-income market to ensure a return on investment, and when the potential market for the new vaccine is limited to low- and middle-income countries, there is an almost unsurmountable valley of death unless philanthropic and public funding intervene<sup>2</sup>.

The needs and unique challenges of vaccines against epidemic pathogens demand innovation in product development pathways. The Merck recombinant vesicular stomatitis virus-Zaire Ebola virus (rVSV-ZEBOV) vaccine was deployed on a large scale during the recent Ebola outbreak in eastern Democratic Republic of the Congo before the product was licensed-even for indications for which no efficacy data were available such as primary prevention in healthcare workers. A second experimental vaccine, Ad26.ZEBOV/MVA-BN-Filo, is now also deployed for the same outbreak and in Rwanda55. Well-informed country leadership and transparent governance of such use are crucial, as is genuine community involvement. The 'animal efficacy rule' that applies when efficacy trials in humans are not feasible or ethical<sup>56</sup> should also be considered for vaccines against epidemic pathogens. The development of Ebola vaccines has shown how this type of 'learning by doing' model can offer early access in humanitarian situations55,57, although it should be stressed that nearly five years after the first Ebola vaccine clinical trials in West Africa, no Ebola vaccine is licensed despite well-documented immunogenicity, safety, and human and/or non-human primate efficacy data. When a crisis such as Ebola is no longer the headline news, the sense of urgency is lost, and regulators and normative committees go back to often extraordinarily long processes.

After a successful phase 3 trial, there is a complex path to the licensing of any new vaccine, which requires reproducibility and safety tests of several batches of vaccines, while manufacturing facilities are finalized. Many countries still request clinical trial data conducted locally, delaying country licensing and implementation considerably, while further raising the costs of development. In Europe, there is advanced harmonization in the regulatory approval of vaccines through the European Medicines Agency (EMA), and in sub-Saharan Africa, the Africa Vaccine Regulatory Forum (AVAREF) is aiming to strengthen regulatory capacity for clinical trials and harmonization of regulatory practices<sup>58</sup>.

Following all of these activities, which can take as long as ten years or more, a new vaccine is now ready for deployment, but a third hurdle can occur between the licensing of a vaccine and broad-scale implementation, which is dependent on both a policy recommendation and the ability to implement. Many years can go by before important new vaccines reach communities in need, the cost of which is measured in human lives that could have been saved as well as money for their development.

There are many contributors to this third hurdle: first is cost, which is especially relevant for countries that are neither wealthy enough to procure vaccines at high cost nor poor enough to receive funding assistance from Gavi, the Vaccine Alliance. However, when a Gavi-eligible country transitions out of the programme owing to an increase in its gross national income per capita, it needs to increasingly mobilize domestic resources or other development assistance<sup>59</sup>. Even when the broader value proposition of a new vaccine is substantial, there remains the question of affordability. Second is the question of country capacity to take on new vaccines; the past decade has been a remarkable era for vaccine introduction, with 113 countries having introduced at least one new vaccine, which represents a real success story<sup>60</sup>. Country capacity to introduce and sustain ever growing programmes involves human and financial resources, and time to build political support and community demand. Both the pneumococcal conjugate and the rotavirus vaccines now have coverage in low-income Gavi countries that meets or exceeds the global average; however, this reflects the fact that not all countries in any income strata have yet introduced these vaccines in spite of their availability<sup>61</sup>. Even high-income countries can experience delays. Thus, in the United Kingdom, a meningococcal B vaccine was licensed in January 2013, recommended for introduction in March 2014, and finally announced for introduction in May 2015. It then took more than 12 months to resolve procurement discussions to enable implementation<sup>62</sup>.

For products that address priority diseases for low-income countries, the uncertainty of the market may risk products collapsing unless a full end-to-end product solution is articulated, with non-commercial support. Inclusion of the new vaccine in the WHO's pre-qualification list is a requirement for procurement through funders such as UNICEF and Gavi. Some of these are vaccines against parasitic diseases, which are much more complex than bacterial or viral vaccines owing to the wide range of antigens with often a complex life cycle that exhibit different antigens relevant for vaccine protection. Thus, the RTS,S vaccine-the first ever malaria vaccine used in a routine immunization system<sup>63</sup>—took nearly 30 years since its creation by GlaxoSmithKline in 1987<sup>64</sup> before the EMA issued a positive scientific opinion in 2015, and the WHO recommended large-scale pilot programmes in 2016. These programmes took another three years to start in several African countries, and demonstrate the sometimes incredibly long development, licensing, and introduction times. The RTS, S malaria vaccine is also an example of a vaccine for which the clinical trial performance of partial protection led to a policy decision to advance in a step-wise manner rather than full programmatic deployment. This may become a more common pathway for future products, in part because these vaccines have performance and implementation characteristics that are more complex than those of current vaccines.

We are entering an era in which the path from vaccine licensing to routine implementation requires more than safety and efficacy data. Policy recommendations for new vaccines may only be realized after implementation research to determine how to ensure use and impact most effectively. Deliberations about cost effectiveness, the full value of vaccine assessments, and country priorities in the face of constrained resources remain drivers for delays associated with the third hurdle. National Immunization Technical Advisory Groups (NITAGS) will be increasingly important to guide evidence-based decision making.

Even after the lengthy and costly trajectory to introduce a new vaccine, ensuring sustainable impact faces a fourth set of hurdles that need to be overcome. These include supply and demand sustainability, and resilience and acceptance of immunization. Logistical issues such as the in-country 'cold chain' system of transporting and storing vaccines at recommended temperatures, procurement management, and the organization of vaccination clinics in remote areas, vaccine hesitancy, and equity of access can all present challenges. In addition, the misuse of vaccination campaigns as political tools has seriously damaged vaccine confidence in areas such as the Philippines, Nigeria, Afghanistan, Italy and Pakistan<sup>65</sup>. Some side effects or limitations of duration of protection may only become obvious after larger scale use, such as for live oral rotavirus vaccination in high-mortality settings<sup>66</sup>, pertussis vaccine<sup>67</sup> and others<sup>68</sup>. A recent example is the results from a retrospective analysis of

#### Review

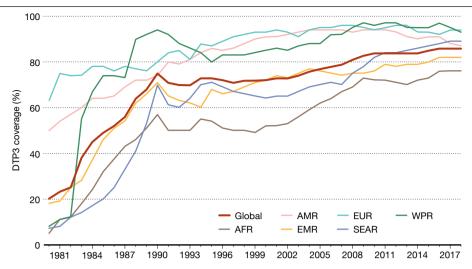


Fig. 1 | Coverage of DTP3 immunization over time globally, combining coverage and regional variations, 1980–2018. The coverage of DTP3 (containing products) immunization improved rapid in the 1980s, with large regional variations. Stagnation over the past 10 years has meant that 19.4 million children remained under-vaccinated or unvaccinated. The thick red line denotes

global coverage, and solid lines represent regional coverages. DTP3, diphtheria, tetanus and pertussis; AFR, AMR, EMR, EUR, SEAR and WPR are WHO subregions of Africa, Americas, Eastern Mediterranean, Europe, South-East Asia and Western Pacific, respectively. Adapted from ref. <sup>76</sup>. Source: WHO/UNICEF coverage estimates 2018 revision, July 2019.

long-term efficacy trials that show that although there is a clear overall population benefit of the Dengvaxia vaccine against dengue, the vaccine also caused an excessive risk of severe dengue in seronegative vaccinees (that is, those not exposed to dengue virus  $^{69}$ ). In the Philippines, this new risk was reported after more than 800,000 school children were vaccinated, prompting a marked reaction by the public in  $2018^{70}$ .

Stock-out events and vaccine manufacturing capacity have been problematic for particular vaccines, even in high-income countries. Manufacturers emphasize the time needed to build and commission a factory<sup>71</sup>. Although manufacturers in middle-income country are now supplying most low-cost vaccines globally, they face low profit margins, ferocious tenders, and often unpredictable procurement schemes. More efficient and modular production technologies may enable decentralized production with lower capital costs.

Each of the four hurdles can be overcome, although the fourth one should be a continuing concern for every national immunization programme. Depending on the phase, they may require different sets of policy actors, and are sometimes a matter of policy, management and leadership, rather than money.

Throughout the development and use of vaccines, vaccine safety is an overriding concern, and requires a continuous and careful scientific and societal assessment. Safety monitoring during manufacturing typically occupies a major part of the process and costs of a vaccine, and is a key element of any vaccine programme. In specific high-income populations, such as in the elderly, personalized medicine approaches have been proposed to maximize both immunogenicity and safety in the presence of chronic conditions and changes related to older age, but large-scale applicability is still questionable at present 72-74.

#### Persistent unmet needs for vaccination

The extraordinary achievement of vaccines is reflected in countries having vaccinated more than 116 million infants in 2018 alone<sup>75</sup>—which represents the largest number ever—and a comparable number of infants were also estimated to have been vaccinated in 2017. The global and regional coverage of diphtheria–tetanus–pertussis (DPT3) vaccination between 1980 and 2018 in Fig. 1 shows overall high coverage with regional variations, but also some stagnation in coverage over the past 10 years<sup>76</sup>. Despite the high coverage, there still remained 19.4 million under-vaccinated or unvaccinated children, who were vulnerable to diseases that they could and should have been protected from.

Substantial improvements in coverage have been achieved in some countries, whereas coverage is regressing in others, often because of social disruption, conflict, or political upheaval, which highlights the extremely dynamic nature of vaccine programme performance.

Around 60% of all children who did not receive basic immunization in 2018 live in ten countries: Angola, Brazil, the Democratic Republic of the Congo, Ethiopia, India, Indonesia, Nigeria, Pakistan, the Philippines and Vietnam<sup>77</sup>. To achieve rapid change in this situation requires the full commitment of governments, supported by international organizations.The Gavi Alliance provides funding for vaccination programmes in lowand low-to middle-income countries, and has had substantial impact. The technical support provided by Gavi partners will be essential to address persistent gaps in vaccine coverage. Consistently delivering vaccines with high coverage, reaching at least the minimum coverage required to achieve herd immunity in line with the basic reproductive ratio of an infection as mentioned above, remains a struggle in many other countries including in middle- and high-income settings, with poor children not being reached<sup>78,79</sup>. For example, in 2017 in the United States, 100,000 children under the age of two (1.3% of the population of that age) were not immunized against DPT and MMR (measles, mumps and rubella), which represents a fourfold increase since 2001<sup>79,80</sup>.

Of particular concern are countries in which vaccination coverage has declined. There are 19 countries that had more than 80% coverage for first-dose measles at some point between 2011 and 2017, but with coverage in 2018 at least 10% lower than their peak coverage. The measles vaccine coverage of those 19 regressing countries now ranges from 38% to 88%, with 10 countries with well below 80% coverage of the regression on vaccine coverage may represent improvements in data rather than actual slippage in coverage. The data systems to monitor both the number of children born and the number of children vaccinated accurately are highly variable in quality \$81,82\$. In some settings, management and reward systems probably incentivize inaccurate reporting of coverage data to meet targets, rather than incentivizing accurate reporting.

Outbreaks of measles, diphtheria and yellow fever are the result of what happens when the world is complacent and immunization coverage declines. Diphtheria outbreaks surged in Russia in the early 1990s; outbreaks of meningitis occurred among Rohingya refugees from Myanmar in refugee camps in 2017; and the transmission of polio persists in parts of Afghanistan and Pakistan<sup>83</sup>. Measles outbreaks are occurring in all regions of the world. The recent 80-fold increase in reported measles

cases in the WHO European Region over four years to more than 82,000 cases in 2018 with 72 deaths 84,85 is a result of a mixture of vaccine refusals, cultural beliefs, and access issues that include interruptions in vaccine supply, such as in Ukraine<sup>86</sup>, and have led to a WHO declaration of a grade 2 health emergency<sup>87</sup>. In the Americas, thousands of cases have been reported in Venezuela owing to the political and economic crisis, with cases also appearing in Brazil, Colombia and Ecuador, and four countries in the WHO European Region (United Kingdom, Albania, Greece and the Czech Republic) have now lost their measles elimination status. The United States is also at risk of losing their measles elimination status.

These outbreaks reflect failures to achieve and maintain high vaccination coverage, community by community. Low vaccination coverage and high heterogeneity in coverage are most deeply seen among African countries where routine rates of immunization in many countries are well below the GVAP targets88.

Since 2010, routine immunization levels have either stagnated or decreased in 54 out of 85 middle-income countries, who do not qualify for support from the Gavi Alliance<sup>78</sup>. Vaccine expenditures per child are often lower in middle-income countries than in low-income Gavi countries. The issue may not be solely due to a lack of funding capability, but may also arise owing to a lack of prioritization of immunization, countries not participating in pooled procurement mechanisms such as via UNICEF, low volumes of vaccines, insufficient efforts to reach vulnerable populations, vaccine choices, and duplicative local regulatory requirements that delay the introduction of new vaccines.

Another unmet need concerns the introduction of new vaccines. Rapid progress has been made to scale up the introduction of vaccines through Gavi investments in low-income countries, but not all vaccines have progressed at the same rapid pace. The adolescent HPV vaccine has been particularly slow to be introduced outside of high-income settings because of programmatic challenges, public-access issues, supply constraints and pricing issues.

Addressing these unmet needs will require persistent implementation of strategies that have been shown to be effective—such as detailed microplanning of local efforts to assure all children are identified and immunized—and special campaigns and approaches such as drone delivery of vaccines in areas that are harder to reach<sup>89</sup>. Systematic evaluation and implementation research should be part of these efforts to develop a firm evidence base for overcoming such programmatic challenges. The WHO has elaborated guidance on implementing high impact immunization programmes (Global Routine Immunization Strategies and Practices. GRISP) to address these unmet needs. Middle-income countries that do not benefit from funding from the Gavi Alliance need procurement mechanisms that can secure more predictable tiered pricing. No set of strategies, however, will succeed without substantially enhanced domestic investment and local political commitment, which continue to limit progress in many parts of the world. As demand for services from communities increases, responsiveness to that demand from governments, the funder of such services in most countries, is more likely<sup>90</sup>.

In addition to the unmet needs related to existing vaccines, nearly half of all deaths from infectious diseases are caused by infections for which no vaccine is available (for example, more than 0.5 million deaths globally in children under 5 years from enteric infections for which there is no vaccine<sup>91</sup>). These should be the priorities for vaccine research and development, as well as improvements needed for particular vaccines such as those against rotavirus, pertussis, polio and yellow fever. Innovations in delivery devices are also important (for example, micropatches, temperature-stable vaccines, improved cold-chain equipment).

#### The equity imperative

Equity has been a primary goal of immunization programmes. To reach those who are in greatest need means addressing issues of vaccine availability, affordability, accessibility, acceptability and financing. An effective immunization system that delivers vaccines with high equity across social and ethnic strata, maternal and community education, and geographies, is a purpose-built programme to deliver impact, and has been shown to be the crucial programmatic target.

Country-level vaccine coverage values mask subnational inequity, risking disease outbreaks and backsliding on achievements of vaccination. Immunization improvements should focus at the subnational level, as well as on other determinants of inequity, not all of which would be addressed by focused supplementary vaccine campaigns.

There is a special case for vaccine development for pathogens that cause epidemics. These diseases have little to no market incentive to drive product development, hence the need for innovative arrangements such as the CEPI<sup>53</sup>, US Biomedical Advanced Research and Development Authority (BARDA)<sup>92</sup> and the European Innovative Medicines Initiative (IMI: https://www.imi.europa.eu)<sup>93</sup>.

Humanitarian crises are another increasing impediment to immunization. The number, size and duration of conflicts, the migration of refugees, and natural disasters have all caused major disruptions to immunization programmes and resulted in serious disease outbreaks. The persisting hurdles to the eradication of polio reveal how political, social and conflict situations can disrupt access to populations and risk violence targeting vaccinators such as in Pakistan and Afghanistan<sup>94</sup>. Nearly 100 polio vaccinators and their security guards have been targeted and killed while attempting to reach children for vaccination 95.

#### The growing challenge of vaccine confidence

Despite the success and wide acceptance of the importance of immunization, there are growing groups of people who delay or refuse vaccines. In 2013, the WHO Strategic Advisory Group of Experts (SAGE) established a working group to investigate the scope and scale of vaccine hesitancy<sup>96</sup>, the US National Vaccine Advisory Committee (NVAC) put together a Vac $cine \, Confidence \, Working \, Group \, to \, investigate \, the \, situation \, in \, the \, United \,$ States (National Vaccine Advisory Committee, 2015), and the European Centre for Disease Prevention and Control (ECDC) published a review of the state of vaccine hesitancy in Europe<sup>97</sup>. In January 2019, the WHO named vaccine hesitancy as one of the top ten global health threats.

Since 2015, the Vaccine Confidence Index (VCI) has surveyed more than 300,000 respondents globally to detect early signals of waning public confidence in vaccine importance, safety and effectiveness, to prompt early intervention where needed (see Fig. 2 for world map of confidence in vaccine safety in 2018). The European Commission adopted the VCI as part of an effort in 2018 to strengthen cooperation against vaccine-preventable diseases98, and the Wellcome Trust used the VCI as part of their 144-country study into public confidence in vaccines (Wellcome Global Monitor 2018)99. Safety was identified as a key issue in both the 2018 European study and the Wellcome report, with public confidence in vaccine safety being consistently lower than the confidence in vaccine effectiveness and importance<sup>99</sup>.

Although a lack of familiarity by both physicians and parents with many childhood diseases because of years of successful vaccination programmes may have a role in a lack of interest in vaccines, the reasons for a decline in vaccine confidence are far more complex. Newer challenges to vaccine confidence include social media campaigns that have disrupted MMR vaccination efforts in southern India, collapsed HPV vaccination efforts in Japan, provoked false scares of vaccine poisoning in Pakistan, and undermined vaccination programmes in Indonesia.

Vaccine confidence issues are highly varied by setting and vaccine. In a three-year review (2015–2017) of the WHO/UNICEF Joint Reporting Form (JRF) completed annually by national immunization programmes, over 90% of the 194 countries reported that they experienced vaccine hesitancy. The top three reasons for hesitancy were: (1) 'risk-benefit (scientific evidence)'-that is, safety concerns; (2) lack of knowledge on the benefits of immunization; and (3) religion, culture and socioeconomic issues100.

#### Review

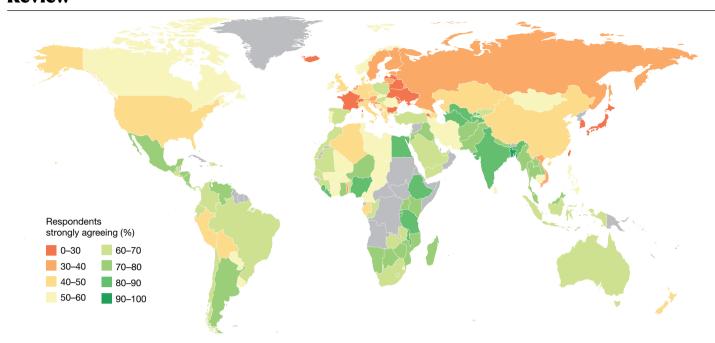


Fig. 2 | Global confidence in vaccine safety in 2018. Levels of confidence in vaccine safety varied considerably across countries and regions, with several countries showing very low levels of confidence. The colour chart at the bottom

shows increasing levels of confidence. Note that the question asked in the survey was 'Do you agree with the following statement: vaccines are safe?'. Source: ref. 99. Map credit: Alexandre De Figueiredo, The Vaccine Confidence Project.

Challenges around building confidence in vaccine safety are well beyond communication, although more accessible public communication around the complex issues of safety and risk benefit analysis are important. What needs to be addressed is not only better communication around the known, albeit sometimes misinterpreted, risks and benefits of vaccination, but also investing in more research in the areas in which the public is asking questions and the science is incomplete. Findings that the ASO3-adjuvanted influenza vaccine Pandemrix was linked to increased cases of narcolepsy in Europe prompted further research, but a systematic review concluded that more research is needed<sup>101</sup>.

Although uncertainty is the norm in science, the political and social worlds of the public have become less tolerant of ambiguity and risk $^{102}$ . New modes of listening to the public, with rapidly evolving technologies to monitor social media, can collect emerging safety questions as well as detect signals of possible issues that need investigation. Working towards better aligned public questions and accessible, evidence-based answers should be a goal. The WHO Vaccine Safety Net initiative is an important resource and can be further built on to address new questions as they emerge, as well as to make new research accessible 103.

Social and political contexts and the reliability of health services are important levers of trust, and a low trust setting will have less tolerance for risk than one with high trust. A 2015 study showed that high trust in immunization services clearly correlated with lower rates of vaccine  $he sitancy ^{104}. \ The \ public's \ experience \ with \ he alth \ services \ and \ he alth$ workers is highly influential in vaccine decision, but both are needed. The Wellcome Global Monitor report showed that in Japan, for example, despite low trust in vaccines and low trust in government, confidence in health providers remained high.

Introducing new vaccines into populations requires adequate time to train and prepare front-line health workers and vaccinators to be ready to manage public questions, and continuing dialogue between scientists and the public will be important to build confidence from the start, as well as to anticipate and manage adverse events.

As mentioned above, reported risks of a recently introduced dengue vaccine<sup>105</sup> in the Philippines amplified into public outrage mediated through Facebook pages, and were made more complex because the events occurred during political elections. The result was a marked drop in public confidence in vaccines more generally from 99.5% in 2015 to 76.2% in 2018, and confidence in vaccine safety plummeted from 99.5% to 65.2%<sup>65</sup> (Fig. 3). The overall drop in public trust affected willingness to accept even the measles vaccine, prompting measles outbreaks with more than 25,000 measles cases and 355 deaths by March 2019<sup>106</sup> and requiring considerable efforts to rebuild public confidence and increase vaccine uptake.

Conflict situations also affect confidence in vaccines and vaccinators owing to an environment of distrust and uncertainty, such as in Pakistan and Afghanistan, and in the Democratic Republic of the Congo, where local violence and conflict in the Ebola-affected areas has been an obstacle to vaccination efforts.

#### The future of immunization

The contribution of immunization to human health, security and prosperity has been matched by few other activities in health and development, and has been crucial for progress in child survival. As immunization coverage among adults is generally low, it is another area in which greater advances can be made.

Addressing the following issues will be crucial to ensure that the effect of vaccination is optimized.

(1) Leadership and funding. Achieving immunization for all those in need should be a top priority for every country. This will require stronger political leadership and a continuing increase in investments in immunization, both domestically and internationally<sup>6</sup>. The power of immunization to achieve wider health and societal benefits should be further documented. The prioritization of vaccines is particularly crucial for middle-income countries that no longer benefit from support from the Gavi Alliance and for countries that are transitioning out of Gavi support.

A successful replenishment of Gavi resources in 2020 for the proposed Gavi 5.0 strategy<sup>107</sup> is vital for the next decade of progress in child survival, and will be a test of the commitment of the international community to immunization and global health.

(2) Universal vaccine coverage and equity. Overcoming the stagnation in reaching all people in need with even the basic vaccines is an overriding

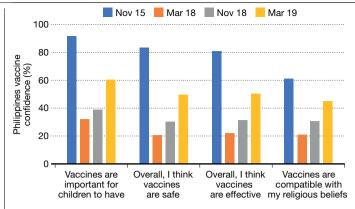


Fig. 3 | Changing levels of vaccine confidence in the Philippines between 2015 and 2019. A marked reduction in public confidence in vaccine safety in the Philippines was partly due to the impact of social media and local politics. Source: The Vaccine Confidence Project<sup>65</sup>, data collected by the Gallup International Association (PSRC).

priority in all countries, especially in those with the lowest coverage and the greatest number of unvaccinated children. As we look towards the next decade, ensuring that vulnerable people in all countries are not left behind should be a top concern, particularly in middle-income countries, as there will be more poor people living there than in poorer countries78.

Ensuring a sustainable and affordable supply of quality vaccines, with differential pricing according to the wealth of a country, is fundamental to achieving sustainability and equity of immunization. Only a few multinational companies are producing vaccines, and a growing number of middleincome manufacturers are major suppliers. There is a risk that continuous lowering of prices may lead to new monopolies, and possibly to higher prices. Healthy vaccine markets with sustainable supply are an important objective for vaccine programmes. Harmonization and strengthening of regulatory capabilities of low- and middle-income countries are essential. Initiatives such as the AVAREF<sup>58</sup> deserve support. The fact that some countries require local clinical trials despite WHO pre-qualification can be a source of major delays in the introduction of vaccines.

(3) People-centred programmes. Immunization programmes can become more effective with a systems-driven and 'precision public health' approach, taking into account local variation in immunization levels, specific needs, cultural specifics, and circumstances of vulnerable populations. Quality data at administrative levels closer to communities should be collected to inform 'micro-planning' and adaptive programme delivery. Innovative efforts such as thoughtful integration of immunization into health services, education systems and elderly care are needed.

As most vaccines have incomplete efficacy, tailored approaches to optimize their impact will be needed, particularly for vaccines against malaria, influenza, dengue and probably HIV when it becomes available.

(4) Vaccine confidence. Vaccine confidence needs to be addressed up front and be an integral part of immunization programmes. Many approaches to increasing vaccine uptake do not take into account the social, historical and political realities of the public for whom information alone is not the antidote to vaccine reluctance. Instead of older demand-creation models, a new model and language of engaging with the public is needed, starting with better listening and prompt responding to concerns as well as building on local capacities. Inclusion of non-traditional partners, new modes of digital communication, social scientists, and religious and traditional leaders have been invaluable in addressing hesitancy around polio vaccination, and the engagement of teenage girls in co-designing social media outreach to address HPV vaccination concerns had positive effects on vaccine uptake in Denmark. With safety anxieties being reported as one of the top reasons for vaccine hesitancy, aligning vaccine safety research with dominant safety concerns will also be important for confidence building.

(5) Investment in research and innovation. Many issues mentioned in the other recommendations require further research in a wide range of disciplines. Product innovation as a result of the formidable progress in immunology and infection pathogenesis has been a strong driver of immunization programmes. There is reluctance of industry to develop vaccines when market incentives are limited, and licensing is uncertain. Although companies such as Merck and Johnson & Johnson invested considerably in the development of candidate Ebola vaccines, partly supported by public funds in North America and Europe, but without a prospect of a return on investment, it would be unrealistic to expect that industry will follow this example for each new emerging pathogen. There is a major role for the public sector and philanthropy to support mechanisms such as the CEPI to develop vaccines for low-income countries<sup>2</sup>. As discussed under the 'second hurdle' on the challenge to fund and conduct late clinical development through to the market introduction for vaccines for which there is no market incentive, there is an urgent need to address this gap, possibly via a specific global initiative or at least a concerted action of several funders. There is also a need for innovation in trial design (for faster trials with smaller sample sizes, and including collection of valuable biosamples to inform correlates of protection) and in trial analysis, as well as in vaccine delivery. Escalating antimicrobial resistance is a powerful incentive to develop vaccines against bacterial infections, malaria, tuberculosis and HIV  $infection {}^{108-110}. Innovation in the delivery of vaccination programmes \\$ is as important as product innovation.

The world cannot afford to turn the clock back on immunization, and ever more innovative vaccines will offer additional opportunities to reduce mortality and improve the quality of life for every person on the planet. This will require the best of science, entrepreneurship, programme implementation on the ground, and politics.

- World Health Organization. Vaccines and Diseases https://www.who.int/immunization/ diseases/en/ (2019)
- Kaslow. D. C. et al. Vaccine candidates for poor nations are going to waste. Nature 564, 337-339 (2018)
- Centers for Disease Control and Prevention (CDC). Ten great public health achievements-United States, 1900-1999. MMWR Morb. Mortal. Wklv. Rep. 48. 241-243 (1999)
- Rappuoli, R., Black, S. & Bloom, D. E. Vaccines and global health; in search of a sustainable model for vaccine development and delivery. Sci. Transl. Med. 11, eaaw2888 (2019). The authors demonstrate how promising and much needed vaccines for global health fail to reach full development because of inadequate financial incentives, and propose mechanisms to overcome this problem.
- Rappuoli, R., Mandl, C. W., Black, S. & De Gregorio, E. Vaccines for the twenty-first century society. Nat. Rev. Immunol. 11, 865-872 (2011).
- World Health Organization. Global Vaccine Action Plan 2011-2020 https://www.who.int/ immunization/global\_vaccine\_action\_plan/GVAP\_doc\_2011\_2020/en/ (2013).
  - This is the global reference for most national immunization programmes
- World Health Organization. State of the World's Vaccines and Immunization. 3rd edition https://www.who.int/immunization/sowvi/en/ (2009)
- Orenstein, W. A. & Ahmed, R. Simply put: vaccination saves lives. Proc. Natl Acad. Sci. USA 114, 4031-4033 (2017).
- Bill & Melinda Gates Foundation. The Goalkeepers Report https://www.gatesfoundation. org/goalkeepers/report (2018).
- Hammitt, L. L. et al. Population effect of 10-valent pneumococcal conjugate vaccine on nasopharyngeal carriage of Streptococcus pneumoniae and non-typeable Haemophilus influenzae in Kilifi. Kenya: findings from cross-sectional carriage studies. Lancet Glob Health 2, e397-e405 (2014)
- World Health Organization, Rubella vaccines: WHO position paper, Wklv. Epidemiol, Rec. 86, 301-316 (2011).
- World Health Organization. Summary of WHO Position Paper on Measles (2017)
- Chan, J. et al. Determining the pneumococcal conjugate vaccine coverage required for indirect protection against vaccine-type pneumococcal carriage in low and middleincome countries: a protocol for a prospective observational study. BMJ Open 8, e021512 (2018).
- Vynnycky, E. & White, R. An Introduction to Infectious Disease Modelling (OUP, 2010).
- Guerra, F. M. et al. The basic reproduction number (R<sub>0</sub>) of measles: a systematic review. Lancet Infect. Dis. 17, e420-e428 (2017).
- Aaby, P., Kollmann, T. R. & Benn, C. S. Nonspecific effects of neonatal and infant vaccination: public-health, immunological and conceptual challenges. Nat. Immunol. 15, 895-899 (2014)
- World Health Organziation. Meeting of the Strategic Advisory Group of Experts on immunization, April 2014 - conclusions and recommendations. Wkly. Epidemiol. Rec. 89,
- Wilder-Smith, A. et al. The public health value of vaccines beyond efficacy: methods, measures and outcomes. BMC Med. 15, 138 (2017).

#### **Review**

- Bärnighausen, T., Bloom, D. E., Cafiero-Fonseca, E. T. & O'Brien, J. C. Valuing vaccination. Proc. Natl Acad. Sci. USA 111, 12313–12319 (2014).
- Ozawa, S. et al. Return on investment from childhood immunization in low- and middleincome countries, 2011-20. Health Aff. (Millwood) 35, 199–207 (2016).
  - The economic analysis in this paper of 10 vaccines administered in 73 low- and middle-income countries between 2011 and 2020 shows that a total of 20 million lives and over \$350 billion could be saved, making vaccines a best buy in public health.
- Brenzel, L. What have we learned on costs and financing of routine immunization from the comprehensive multi-year plans in GAVI eligible countries? Vaccine 33 (Suppl. 1), A93–A98 (2015).
- Zhou, F. et al. Economic evaluation of the routine childhood immunization program in the United States, 2009. Pediatrics 133, 577-585 (2014).
- Riumallo-Herl, C. et al. Poverty reduction and equity benefits of introducing or scaling up measles, rotavirus and pneumococcal vaccines in low-income and middle-income countries: a modelling study. BMJ Global Health 3, e000613 (2018).
- Chang, A. Y. et al. The equity impact vaccines may have on averting deaths and medical impoverishment in developing countries. *Health Aff. (Millwood)* 37, 316–324 (2018).
- Jit, M. et al. The broader economic impact of vaccination: reviewing and appraising the strength of evidence. BMC Med. 13, 209 (2015).
  - This paper considers the broader economic benefits around childhood development, household behaviour, and macro-economic indicators in addition to the usual microeconomic evaluations of immunization.
- Smith, R. D., Keogh-Brown, M. R., Barnett, T. & Tait, J. The economy-wide impact of pandemic influenza on the UK: a computable general equilibrium modelling experiment. Br. Med. J. 339, b4571 (2009).
- Putri, W. C. W. S., Muscatello, D. J., Stockwell, M. S. & Newall, A. T. Economic burden of seasonal influenza in the United States. Vaccine 36, 3960–3966 (2018).
- The World Bank. From Panic and Neglect to Investing in Health Security: Financing Pandemic Preparedness at a National Level https://www.worldbank.org/en/topic/ pandemics/publication/from-panic-neglect-to-investing-in-health-security-financing-pandemic-preparedness-at-a-national-level (2017).
- VoICE. The Value of Immunization Compendium of Evidence https://immunizationevidence. org/about/ (2019).
- Grassly, N. C., Kang, G. & Kampmann, B. Biological challenges to effective vaccines in the developing world. *Phil. Trans. R. Soc. Lond. B* 370, 20140138 (2015).
  - This article introduces a series of papers presented at a scientific meeting held at the Royal Society in London in 2015 that address the wide range of biological reasons for the variability in vaccine efficacy, such as age, sex, environment, genetics and coinfections
- Permar, S., Levy, O., Kollman, T. R., Singh, A. & De Paris, K. Early life HIV-1 immunization: providing a window for protection before sexual debut. AIDS Res. Hum. Retroviruses 34, 823–827 (2018).
- Lee, A. H. et al. Dynamic molecular changes during the first week of human life follow a robust developmental trajectory. Nat. Commun. 10, 1092 (2019).
- Kobayashi, M. et al. Group B Streptococcus vaccine development: present status and future considerations, with emphasis on perspectives for low and middle income countries. F1000 Res. 5, 2355 (2016).
- Madhi, S. A. & Dangor, Z. Prospects for preventing infant invasive GBS disease through maternal vaccination. Vaccine 35, 4457–4460 (2017).
- 35. Yen, L. M. & Thwaites, C. L. Tetanus. Lancet 393, 1657-1668 (2019).
- Munoz, F. M. et al. The Fourth International Neonatal and Maternal Immunization Symposium (INMIS 2017): Toward integrating maternal and infant immunization programs. mSphere 3, e00221-18 (2018).
- Westphal, D. W. et al. A protracted mumps outbreak in Western Australia despite high vaccine coverage: a population-based surveillance study. Lancet Infect. Dis. 19, 177-184 (2019)
- Fields, V. S. et al. Mumps in a highly vaccinated Marshallese community in Arkansas, USA: an outbreak report. Lancet Infect. Dis. 19, 185–192 (2019).
- Hotez, P. America and Europe's new normal: the return of vaccine-preventable diseases. Pediatr. Res. 85, 912–914 (2019).
- Alliance for Aging Research. The Silver Book®: Infectious Diseases and Prevention through Vaccination http://www.silverbook.org/publication/infectious-diseases/ (2013).
- de Gomensoro, E., Del Giudice, G. & Doherty, T. M. Challenges in adult vaccination. Ann. Med. 50, 181–192 (2018).
- International Federation on Ageing. World Coalition on Adult Vaccination https://www.ifafiv.org/project/adult-immunization-advocacy-2/ (2019).
- Kim, D. K., Riley, L. E., Harriman, K. H., Hunter, P. & Bridges, C. B. Recommended immunization schedule for adults aged 19 years or older, United States, 2017. Ann. Intern. Med. 166, 209–219 (2017).
- National Health Service. Vaccinations; Vaccines Given to Adults https://www.nhs.uk/ conditions/yaccinations/?tabname=adults (accessed 1 September 2019).
- Bridges, C. B. et al. Meeting the challenges of immunizing adults. Vaccine 33 (Suppl. 4), D114–D120 (2015).
- Phillips, D. E., Dieleman, J. L., Lim, S. S. & Shearer, J. Determinants of effective vaccine coverage in low and middle-income countries: a systematic review and interpretive synthesis. BMC Health Serv. Res. 17, 681 (2017).
- Leidner, A. J. et al. Cost-effectiveness of adult vaccinations: a systematic review. Vaccine 37, 226–234 (2018).
  - This paper provides insights into the biological challenges to advance the concept of a life course of vaccination.
- Derhovanessian, E. & Pawelec, G. Vaccination in the elderly. Microb. Biotechnol. 5, 226–232 (2012).
- Public Health England. Complete Routine Immunisation Schedule https://www.gov.uk/ government/publications/the-complete-routine-immunisation-schedule (2019).
- Centers for Disease Control and Prevention. Vaccines & Immunizations https://www.cdc. gov/vaccines/index.html (2016).

- Public Health England. Immunisation of Healthcare and Laboratory Staff: the Green Book, Chapter 12 https://www.gov.uk/government/publications/immunisation-of-healthcareand-laboratory-staff-the-green-book-chapter-12 (2013).
- Sarkar, S., Zlojutro, A., Khan, K. & Gardner, L. Measles resurgence in the USA: how international travel compounds vaccine resistance. *Lancet Infect. Dis.* 19, 684–686 (2019).
- 3. CEPI. The Coalition for Epidemic Preparedness Innovations https://cepi.net/ (2019).
- Butler, D. Translational research: crossing the valley of death. Nature 453, 840–842 (2008).
- Winslow, R. L. et al. Immune responses to novel adenovirus type 26 and modified vaccinia virus Ankara-vectored Ebola vaccines at 1 year. J. Am. Med. Assoc. 317, 1075–1077 (2017).
- Snoy, P. J. Establishing efficacy of human products using animals: the US food and drug administration's "animal rule". Vet. Pathol. 47, 774–778 (2010).
- Henao-Restrepo, A. M. et al. Efficacy and effectiveness of an rVSV-vectored vaccine in preventing Ebola virus disease: final results from the Guinea ring vaccination, open-label, cluster-randomised trial (Ebola Ca Suffit!). Lancet 389, 505–518 (2017).
- Akanmori, B., Bellah, A., Ward, M. & Rago, L. The African vaccine regulatory forum (AVAREF): a platform for collaboration in a public health emergency. WHO Drug Inf. 29, 127–132 (2015).
- Kallenberg, J. et al. Gavi's transition policy: moving from development assistance to domestic financing of immunization programs. Health Aff. (Millwood) 35, 250–258 (2016).
- World Health Organization. Assessment Report of the Global Vaccine Action Plans Strategic Advisory Group of Experts on Immunization https://apps.who.int/iris/handle/10665/276967 (2018).

# This is the annual update on the state of immunization in the world, with a wealth of data.

- World Health Organization. Immunization, Vaccines and Biologicals. Data, Statistics and Graphics https://www.who.int/immunization/monitoring\_surveillance/data/en/ (2019).
- 62. Findlow, J. Vaccines for the prevention of meningococcal capsular group B disease: what have we recently learned? *Hum. Vaccin. Immunother.* **12**, 235–238 (2016).
- World Health Organization. Malaria Vaccine Pilot Launched in Malawi https://www.who. int/news-room/detail/23-04-2019-malaria-vaccine-pilot-launched-in-malawi (2019).
- Roland, D. Malaria vaccine: GSK's thirty-year quest to eradicate a global killer. Telegraph (8 October 2013).
- Larson, H. J., Hartigan-Go, K. & de Figueiredo, A. Vaccine confidence plummets in the Philippines following dengue vaccine scare: why it matters to pandemic preparedness. Hum. Vaccin. Immunother. 15, 625–627 (2019).
- Clark, A. et al. Efficacy of live oral rotavirus vaccines by duration of follow-up: a metaregression of randomised controlled trials. Lancet Infect. Dis. 19, 717–727 (2019).
- Burdin, N., Handy, L. K. & Plotkin, S. A. What is wrong with pertussis vaccine immunity? The problem of waning effectiveness of pertussis vaccines. Cold Spring Harb. Perspect. Biol. 9, a029454 (2017).
- Khan, M. I. et al. Barriers to typhoid fever vaccine access in endemic countries. Res. Rep. Trop. Med. 8, 37-44 (2017)
- Wilder-Smith, A. et al. Deliberations of the Strategic Advisory Group of Experts on Immunization on the use of CYD-TDV dengue vaccine. *Lancet Infect. Dis.* 19, e31–e38 (2019)
- Cabico, G. K. How the Dengvaxia scare helped erode decades of public trust in vaccines. Philstar Global Corp. (6 February 2019).
- 71. Plotkin, S., Robinson, J. M., Cunningham, G., Iqbal, R. & Larsen, S. The complexity and cost of vaccine manufacturing an overview. *Vaccine* **35**, 4064–4071 (2017).

# An important paper to help to understand the complexity and cost of vaccine manufacturing.

- Poland, G. A., Ovsyannikova, I. G. & Jacobson, R. M. Personalized vaccines: the emerging field of vaccinomics. Expert Opin. Biol. Ther. 8, 1659–1667 (2008).
- Poland, G. A., Ovsyannikova, I. G., Jacobson, R. M. & Smith, D. I. Heterogeneity in vaccine immune response: the role of immunogenetics and the emerging field of vaccinomics. *Clin. Pharmacol. Ther.* 82, 653–664 (2007).
- Mentzer, A. J., O'Connor, D., Pollard, A. J. & Hill, A. V. Searching for the human genetic factors standing in the way of universally effective vaccines. *Phil. Trans. R. Soc. Lond. B* 370, 20140341 (2015).
- UNICEF & World Health Organization. Progress and Challenges with Achieving Universal Immunization Coverage https://www.who.int/immunization/monitoring\_surveillance/ who-immuniz.pdf?ua=1 (2019).
- UNICEF & World Health Organization. Progress Towards Global Immunization Goals 2018 https://www.who.int/immunization/monitoring\_surveillance/SlidesGlobalImmunization. pptx?ua=1 (2019).
- World Health Organization. Immunization Coverage https://www.who.int/en/news-room/ fact-sheets/detail/immunization-coverage (2019).
- 78. Berkley, S. Vaccination lags behind in middle-income countries. *Nature* **569**, 309 (2019).
- Hill, H. A., Elam-Evans, L. D., Yankey, D., Singleton, J. A. & Kang, Y. Vaccination coverage among children aged 19-35 months - United States, 2017. MMWR Morb. Mortal. Wkly. Rep. 67, 1123–1128 (2018).
- 80. Zimlich, R. How Many Kids are Completely Unvaccinated? https://www.contemporary pediatrics.com/pediatrics/how-many-kids-are-completely-unvaccinated (2018).
- Gong, W. et al. Comparison of three rapid household survey sampling methods for vaccination coverage assessment in a peri-urban setting in Pakistan. *Int. J. Epidemiol.* 48, 583–595 (2018).
- Cutts, F. T., Izurieta, H. S. & Rhoda, D. A. Measuring coverage in MNCH: design, implementation, and interpretation challenges associated with tracking vaccination coverage using household surveys. PLoS Med. 10, e1001404 (2013).
- Lam, E., McCarthy, A. & Brennan, M. Vaccine-preventable diseases in humanitarian emergencies among refugee and internally-displaced populations. *Hum. Vaccin. Immunother.* 11, 2627–2636 (2015).
- World Health Organization Regional Office for Europe. Measles in Europe: Record Number of Both Sick and Immunized http://www.euro.who.int/en/media-centre/sections/pressreleases/2019/measles-in-europe-record-number-of-both-sick-and-immunized (2019).

- European Centre for Disease Prevention and Control. Monthly Measles and Rubella Monitoring Report, April 2019 https://ecdc.europa.eu/en/publications-data/monthlymeasles-and-rubella-monitoring-report-april-2019 (2019).
- The Lancet. Measles, war, and health-care reforms in Ukraine. Lancet 392, 711 (2018).
- World Health Organization Regional Office for Europe. Over 100,000 People Sick With Measles in 14 Months: with Measles Cases at an Alarming Level in the European Region, WHO Scales up Response http://www.euro.who.int/en/media-centre/sections/press releases/2019/over-100-000-people-sick-with-measles-in-14-months-with-measlescases-at-an-alarming-level-in-the-european-region,-who-scales-up-response (2019).
- Mosser, J. F. et al. Mapping diphtheria-pertussis-tetanus vaccine coverage in Africa, 2000-2016: a spatial and temporal modelling study. Lancet 393, 1843-1855 (2019). This paper shows that the monitoring of vaccine coverage over time, at local rather than national levels, is crucial for programme management, improvement and for implementing operational approaches that enhance vaccine coverage equity.
- Murray, J. Vaccines by air as drone medicine service takes off in Ghana, The Guardian (25 April 2019).
- 90. World Health Organization, Immunization, Vaccines and Biologicals: Global Routine Immunization Strategies and Practices (GRISP) https://www.who.int/immunization/ programmes systems/policies strategies/GRISP/en/ (2019).
- 91. GBD 2017 Causes of Death Collaborators. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017; a systematic analysis for the Global Burden of Disease Study 2017. Lancet 392, 1736-1788 (2018).
- US Department of Health and Human Services. Biomedical Advanced Research and Development Authority https://www.phe.gov/about/barda/Pages/default.aspx (2019).
- 93. Panteli, D. & Edwards, S. Ensuring Access to Medicines: How to Stimulate Innovation to Meet Patients' Needs? (eds Richardson, E. et al.) (World Health Organization Regional Office for Europe, 2018).
- 94. McNeil, D. G. Jr. Polio cases surge in Pakistan and Afghanistan. The New York Times (2019).
- DW NEWS. Pakistani Government Suspends Polio Vaccination Drives https://www.dw.com/ cda/en/pakistani-government-suspends-polio-vaccination-drives/av-48677882 (2019).
- 96. World Health Organization. Report of the SAGE Working Group on Vaccine Hesitancy https://www.who.int/immunization/sage/meetings/2014/october/1\_Report\_WORKING\_ GROUP\_vaccine\_hesitancy\_final.pdf (2014).
- European Centre for Disease Prevention and Control. Rapid Literature Review on Motivating Hesitant Population Groups in Europe to Vaccinate https://ecdc.europa.eu/en/ publications-data/rapid-literature-review-motivating-hesitant-population-groups-europevaccinate (2015).
- 98. European Commission. The State of Vaccine Confidence in the EU: 2018 https://www. vaccineconfidence.org/research/the-state-of-vaccine-confidence-in-the-eu-2018/ (2018).
- The Wellcome Trust. Wellcome Global Monitor 2018 Chapter 5: Attitudes to Vaccines https://wellcome.ac.uk/reports/wellcome-global-monitor/2018/chapter-5-attitudesvaccines (2019).
  - This is one of the largest studies to evaluate of people's attitudes, interest, trust and beliefs about science including immunization.
- 100. Lane, S., MacDonald, N. E., Marti, M. & Dumolard, L. Vaccine hesitancy around the globe: Analysis of three years of WHO/UNICEF Joint Reporting Form data-2015-2017. Vaccine 36, 3861-3867 (2018).
- Cohet, C. et al. Safety of ASO3-adjuvanted influenza vaccines: a review of the evidence. Vaccine 37, 3006-3021 (2019)
- 102. Karafillakis, E. & Larson, H. J. The benefit of the doubt or doubts over benefits? A systematic literature review of perceived risks of vaccines in European populations. Vaccine 35, 4840-4850 (2017)
  - This systematic review of vaccination risk perceptions and concerns published between 2004 and 2014 across Europe shows that the main concern was vaccine

- safety, followed by perceptions of low risk of contracting the vaccine-preventable diseases and, if contracted, that they were not severe
- 103. World Health Organization. Vaccine Safety Net https://www.who.int/vaccine\_safety/ initiative/communication/network/vaccine\_safety\_websites/en (2019).
- 104. London School of Hygiene & Tropical Medicine. The State of Vaccine Confidence 2015 https://www.vaccineconfidence.org/research/2015-vaccine-confidence (2015). This was the first benchmark report by the Vaccine Confidence Project, providing an overview on the key issues affecting confidence globally.
- 105. Sridhar, S. et al. Effect of dengue serostatus on dengue vaccine safety and efficacy. N. Engl. J. Med. 379, 327-340 (2018).
- 106. UNICEF & World Health Organization. Situation Report 8: Measles Outbreak https:// reliefweb.int/report/philippines/unicef-who-philippines-measles-outbreak-situationreport-8-2-april-2019 (2019).
- 107. Gavi. Gavi Board Starts Framing Alliance's Approach to 2021–2025 Period https://www. gavi.org/library/news/press-releases/2018/gavi-board-starts-framing-alliance-sapproach-to-2021-2025-period/(2018).
- 108. Bloom, D. E., Black, S., Salisbury, D. & Rappuoli, R. Antimicrobial resistance and the role of vaccines, Proc. Natl Acad. Sci. USA 115, 12868-12871 (2018).
- 109. Klugman, K. P. & Black, S. Impact of existing vaccines in reducing antibiotic resistance: primary and secondary effects. Proc. Natl Acad. Sci. USA 115, 12896-12901 (2018)
- 110. Sevilla, J. P., Bloom, D. E., Cadarette, D., Jit, M. & Lipsitch, M. Toward economic evaluation of the value of vaccines and other health technologies in addressing AMR. Proc. Natl Acad. Sci. USA 115, 12911-12919 (2018).
- Plotkin, S. L. & Plotkin, S. A. in *Plotkin's Vaccines* 7th edn (eds Plotkin, S. A. et al.) 1–15 (Elsevier, 2018).

Acknowledgements Vaccine research by P.P., H.J.L., B.K. and S.S. is supported by the EU's Innovative Medicines Initiative, the Medical Research Council of UK research and Innovation, The Bill & Melinda Gates Foundation, the European Commission, PATH, the National Institute for Health Research, GSK, Pfizer, and the Wellcome Trust. This Review does not necessarily reflect the position of the World Health Organization.

Author contributions All authors contributed to the development and writing of the manuscript.

Competing interests P.P. is a board member of the Coalition for Epidemic Preparedness Innovation (CEPI), and the Global Health Innovative Technology Fund; H.J.L. has a grant from GSK, and is on an Advisory Board of Takeda: K.L.O. is the director of the Department of  $Immunization, Vaccines, and Biologicals at the World Health Organization; J.N. is {\tt Director}\ of$ the African Centres for Disease Control and Prevention of the African Union, and is a Board member of CEPI. S.S. is the Director General of the Centre for Vaccine Development, Ministry of Health, Mali and a member of the WHO/SAGE Meningitis Working Group, and is an alternate Board Member of Gavi The Vaccine Alliance; B.K. is the Director of the LSHTM Vaccine Centre, and has grants from Pfizer and GSK.

#### Additional information

Correspondence and requests for materials should be addressed to P.P.

Peer review information Nature thanks Alan Barrett, Kathryn Edwards and Rino Rappuoli for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints.

© Springer Nature Limited 2019

# A new twenty-first century science for effective epidemic response

https://doi.org/10.1038/s41586-019-1717-y

Received: 10 June 2019

Accepted: 24 September 2019

Published online: 6 November 2019

Juliet Bedford<sup>1</sup>, Jeremy Farrar<sup>2\*</sup>, Chikwe Ihekweazu<sup>3</sup>, Gagandeep Kang<sup>4</sup>, Marion Koopmans<sup>5</sup> & John Nkengasong<sup>6</sup>

With rapidly changing ecology, urbanization, climate change, increased travel and fragile public health systems, epidemics will become more frequent, more complex and harder to prevent and contain. Here we argue that our concept of epidemics must evolve from crisis response during discrete outbreaks to an integrated cycle of preparation, response and recovery. This is an opportunity to combine knowledge and skills from all over the world—especially at-risk and affected communities. Many disciplines need to be integrated, including not only epidemiology but also social sciences, research and development, diplomacy, logistics and crisis management. This requires a new approach to training tomorrow's leaders in epidemic prevention and response.



When *Nature* published its first issue in 1869<sup>1</sup>, a new understanding of infectious diseases was taking shape. The work of William Farr<sup>2</sup>, Ignaz Semmelweis<sup>3</sup>, Louis-René Villermé<sup>4</sup> and others had been published; John Snow had traced the source of a cholera epidemic in London<sup>5</sup> (although Robert Koch had not yet isolated the bacterium that caused it<sup>6</sup>). The science of epidemiology has described patterns of disease in human populations, investigated the causes of those diseases, evaluated attempts to control them<sup>7</sup> and has been the foundation for public health responses to epidemic infections for over 100 years. Despite great technological progress and expansion of the field, the theories and practices of infectious disease epidemiology are struggling to keep pace with the transitional nature of epidemics in the twenty-first century and the breadth of skills needed to respond to them.

Epidemiological transition theory has focused mostly on the effects of demographic and socioeconomic transitions on well-known preventable infections and a shift from infectious diseases to non-communicable diseases. However, it has become clear that current demographic transitions—driven by population growth, rapid urbanization, deforestation, globalization of travel and trade, climate change and political instability—also have fundamental effects on the dynamics of infectious diseases that are more difficult to predict. The vulnerability of populations to outbreaks of zoonotic diseases such as Ebola, Middle East respiratory syndrome (MERS) and Nipah has increased, the rise and spread of drug-resistant infections, marked shifts in the ecology of known vectors (for example, the expanding range of *Aedes* mosquitoes) and massive amplification of transmission through globally

connected, high-density urban areas (particularly relevant to Ebola, dengue, influenza and severe acute respiratory syndrome-related coronavirus SARS-CoV). These factors and effects combine and interact, fuelling more-complex epidemics.

Although rare compared to those diseases that cause the majority of the burden on population health, the nature of such epidemics disrupts health systems, amplifies mistrust among communities and creates high and long-lasting socioeconomic effects, especially in low- and middle-income countries. Their increasing frequency demands attention. As the Executive Director of the Health Emergencies Program at the World Health Organization (WHO) has said: "We are entering a very new phase of high-impact epidemics... This is a new normal, I don't expect the frequency of these events to reduce."9.

We have to act now but act differently: a broader foundation is required, enhancing traditional epidemiology and public health responses with knowledge and skills from a number of areas (Table 1). Many of these areas have long been associated with epidemic preparedness and response, but they must now stop being seen as esoteric 'nice things to have', and instead become fully integrated into the critical planning and response to epidemics.

This will require considerable changes by the global public health community in the way that we respond to epidemics today and how we prepare for and seek to prevent those of tomorrow. It will mean reshaping the global health architecture of the response to epidemics and transforming how we train new generations of researchers and practitioners for the epidemics of the future<sup>10</sup>.

The modern research culture—often shaped by the behaviour of funders—has required many researchers to specialize in narrow fields, with less emphasis on translation than on field-specific innovations. Although this siloed landscape has brought major advances in global health, it is not fit for the transitional phase of epidemic diseases: rapidly evolving, high-impact events bring together communities, responders and researchers who do not routinely interact. Different assumptions, cultures and practices, each of which may be widely accepted within a

<sup>1</sup>Anthrologica, Oxford, UK. <sup>2</sup>Wellcome, London, UK. <sup>3</sup>Nigeria Centre for Disease Control, Abuja, Nigeria. <sup>4</sup>Translational Health Science and Technology Institute, Faridabad, India. <sup>5</sup>Department of Viroscience, Erasmus University Medical Center, Rotterdam, The Netherlands. <sup>6</sup>Africa Centres for Disease Control and Prevention, African Union, Addis Ababa, Ethiopia. <sup>\*</sup>e-mail: i.farrar@wellcome.ac.uk

Table 1 | Selected key areas to integrate into twenty-first century epidemic responses

Area	Key areas and/or disciplines		
Governance and infrastructure	Local, national and international organizations; integrate accountability and transparency across multiple stakeholders; improve data sharing, improve logistics and crisis management		
Engagement and communication	Encourage a community-led response, community engagement and health diplomacy		
Social sciences	Anthropology, political science, human geography, linguistics		
Ethics	Consent, clinical trial designs		
Emerging technologies	Pathogen genomics, metagenomics, systems serology and analytics, data science and artificial intelligence		
Research and development	Diagnostics, therapeutics and vaccines		
One Health	Ecology and environmental, veterinary and agricultural sciences		

particular community, make working together in outbreak situations more challenging. Fundamental to success is respect and understanding of the contribution each party brings. In a successfully integrated approach, we each have to realize that our knowledge and skills are a small part of a rapidly expanding toolkit (Box 1). We need to understand major trends in research and how and when they may influence the response to an epidemic, develop new research to strengthen the support that we can provide across other areas and learn to operate in multi-stakeholder situations—including, at times, as part of a critical debate to bring better practices to the fore.

Central to this approach must be the communities who are at risk and those affected by epidemics: local people are the first responders to any outbreak and their involvement in the preparation and response activities is essential. From communities, through local and regional health authorities, national public health institutes and international organizations-including many essential partners in sectors beyond public health-the integrated approach must be supported. The WHO, in particular, has a critical part to play, using its unique mandate not to lead every aspect of preparation, response and recovery, but to change its practices, facilitate integration with and among others, and ensure accountabilities are built in from the bottom to the top.

#### Nineteenth and twentieth century epidemiology

A wave of cholera epidemics across Europe in the 1830s and 1840s catalysed a new era of 'infectious disease diplomacy'11 globally. Nations recognized that infections do not stop at borders and that therefore multilateral collaboration is essential to protecting citizens from lethal epidemics. The development of germ theory through the second half of the nineteenth century12 transformed ideas about the causes of infections, informing scientific research as well as clinical responses. Scientific understanding translated into vaccines<sup>13</sup> and antibiotics, while programmes for child health, hygiene, clean water and sanitation became common in the twentieth century. As a result, childhood diseases such as measles and mumps became rare, smallpox was eventually eradicated14 and polio was eliminated from all but a handful of countries<sup>15</sup>. Many people thought that infectious diseases would soon be history. Sir Frank Macfarlane Burnet is often cited for his remark in the 1970s that, with the emergence of new diseases being a distant prospect, "the future of infectious diseases will be very dull"16.

Although the focus in high-income nations turned to non-communicable diseases, which constituted a considerable and increasing burden on the health of their citizens, infectious diseases did not disappear. Some endemic infections such as malaria and tuberculosis

#### Box 1

# Non-traditional tools for epidemics

#### Artificial intelligence

Advances in computer science and computing speeds have led to a number of applications of artificial intelligence across society<sup>84</sup>. Applications in epidemiology include tracking online searches about disease symptoms to aid early detection of epidemics, although more sophisticated methods may be required before artificial intellegence becomes a reliable detection tool<sup>85</sup>.

#### Crystallography

Modern X-ray diffraction and electron microscopy can reveal structures of viruses and antibodies in such detail that it is possible to identify specific sites of vulnerability on the virus. A previous study showed how such techniques identified an antibody that was much more potent against respiratory syncytial virus than the only currently available intervention<sup>86</sup>.

#### Platform vaccine technology

Developing vaccines for emerging infectious diseases has many challenges, including the time it takes, a limited market and strict regulatory requirements for products that will be given to healthy people<sup>87</sup>. Platform technologies use one underlying approach with standardized processes and some antigen-specific optimization to speed up both development and manufacture of vaccines. For example, vector-based platforms combine an antigen, or a gene for an antigenic protein or peptide, in a virus-like particle or liposome. Such platform technologies have the potential to deliver vaccines a few months after an emerging pathogen is identified and sequenced, rather than years<sup>88</sup>.

were not susceptible to elimination strategies, and new diseases with epidemic and pandemic potential emerged. Ebola virus disease was first identified in the 1970s, HIV/AIDS in the 1980s, Nipah virus in the 1990s, SARS and MERS at the start of the twenty-first century, and many more have since been identified. Far from becoming 'very dull'. the field of infectious disease epidemiology has sometimes struggled to adapt: as late as 1990, respected researchers used a nineteenth century 'law' of epidemiology to make predictions about the AIDS epidemic-these turned out to be vast underestimates<sup>17</sup>. Advances in other fields gave epidemiology the chance to evolve. In 2001, when the editors of the *International Journal of Epidemiology* provocatively asked whether it was time to 'call it a day' 18 given the putative power of genomics to explain diseases over the capacity of epidemiologists to describe them, their conclusion was that it had the potential to positively transform epidemiology as much as the rise of germ theory a century earlier.

#### The new normal

At least 150 pathogens that affect humans have been identified as emerging, re-emerging or evolving since the 1980s<sup>19</sup>, while increasing rates of antimicrobial resistance threaten to make formerly controlled infections, such as malaria, untreatable<sup>20</sup>—this also limits our ability to control their epidemic potential. The demographic transition is driving much of this: human society is becoming more urban than rural for the first time in our history, bringing large numbers of people (and often animals) together in densely populated areas<sup>21</sup>. Agricultural and forestry practices are changing the relationships between people, animals and our respective habitats<sup>22</sup>. Travel is more accessible around the world,

#### **Review**

so migration, trade and tourism bring more people into contact and thus affect disease transmission<sup>23</sup>. Climate change has many effects on ecosystems and environments, not least in changing the habitats and migratory habits of disease vectors<sup>24</sup>. States with weak health systems are far less likely to cope with or recover from multiple emergent demands without damaging routine services<sup>25</sup>. Inequalities<sup>26</sup>, inequities and distrust in national structures and institutions compound people's vulnerabilities<sup>27</sup>. Conflict increases the risk of epidemics and makes responding to them close to impossible<sup>28</sup>.

Since 2000, there have been several outbreaks of Ebola (including the two biggest in history), not to mention outbreaks of SARS, MERS, Nipah, influenza A subtype H5N1, yellow fever, Zika and the continued spread of dengue. Epidemics overlap and run into each other, yet the world is not currently equipped to cope with this increasing burden of multiple public health emergencies. Preparing for epidemics, therefore, requires global health, economic and political systems to be integrated just as much as infectious disease epidemiology, translational research and development, and community engagement.

#### Essential areas in epidemic response

#### Governance and infrastructure

Epidemics represent shared risks that cross borders and all of society. Health systems, routine care, trust in governments, travel, trade, business—all are disrupted during an epidemic. With such broad risks, the preparation and response must be nationally owned and led, internationally supported and undertaken with a whole-of-society approach. Some initiatives have started to build frameworks for this to happen in a coordinated way. For example, the WHO's Pandemic Influenza Preparedness Framework brings together nation states, industry, other stakeholders and the WHO to implement a global approach to pandemic preparedness and response<sup>29</sup>.

A focus must be building coordinated regional and country expertise, resources and capacity through national and regional public health institutions<sup>30</sup>. This brings its own challenges—governance of institutions, leadership, collaborations and interventions have to be impeccable or misconduct can thrive<sup>31</sup>. Unwelcome in itself, misuse of funding, resources or people within efforts intended to support an epidemic response will also undermine trust in the organizations that respond to an outbreak and, in turn, prolong the outbreak.

Key governance components include drafting policies in advance and being willing to implement those policies for data collection and sharing during epidemics. They must be flexible enough to enable affected communities and nations to retain ownership of the response, while drawing on international expertise to find the best possible response. Governance should also include processes for vaccine and therapeutic approvals during outbreaks. However, it is clear that the centre of gravity for leadership, governance and implementation must be where the need is greatest if these are to truly deliver.

In 1971, Julian Tudor Hart proposed the inverse care law: "The availability of good medical care tends to vary inversely with the need for it in the population served."32. An analogue of the inverse care law can be applied to public health and epidemiology. Expertise in these fields has traditionally gravitated towards centres of excellence in Europe and the United States. Of course, high-income countries are not immune to the disruption associated with epidemics, especially in an era of misinformation and growing mistrust in authorities and public health initiatives. However, the centre of gravity must shift so that globally representative distributed networks of collaborating centres can jointly ensure coverage in the regions that urgently need these skills on the ground<sup>33</sup>. International collaborations remain important; however, strengthening epidemiology, public health and laboratory capacity in low- and middleincome countries is essential<sup>34</sup>. Collaborative interventions should not be limited to when there is a major outbreak, but be integrated into regular interactions.

Capacity, resources, expertise and governance can be supported by the increasing role for regional and national centres of disease control. The US Centers for Disease Control (CDC) lends its expertise all around the world in addition to protecting the US population. In 2004, the European CDC started, followed by the China CDC in 2015 and by the Africa CDC in 2017. Although more can be done to improve data sharing and access to laboratories, the networks and connections between these centres have strengthened all of their work, as well as having a positive effect on public health systems in low- and middle-income countries.

#### **Engagement and communication**

During the pan-European wave of cholera in the 1830s, there were riots across the continent: doctors, nurses and pharmacists were murdered, hospitals and medical equipment destroyed 27. Similar reports today usually come from communities that have not had positive prior interactions with public health initiatives, and thus the encounter with national or international teams who arrive only in response to a 'new' disease means that trust can never be assumed and has to be earned on both sides. Engagement needs to start before an outbreak—ensuring that patients, their families and their communities are at the centre of all public health is essential for the successful prevention and response to epidemics. There is no public health without the support of the community.

For example, the early detection of disease events will be improved if more national and regional public health institutions establish community event-based surveillance systems. Communities are the first to know when something unusual happens<sup>35</sup>—therefore training and mobilizing community volunteers to report such occurrences is a costeffective way to rapidly detect diseases and contain them at the source. This will also help to sustain engagement between communities and the organizations that respond to outbreaks. Furthermore, improved information flow between the community and the public health system should provide a better understanding of local social networks to complement other means of tracking chains of transmission between individuals and places. This can be the community themselves, or it might be veterinarians who see clusters of sick animals, or nurses and doctors who care for patients in primary care—or it may be teams that are often forgotten in public health initiatives, such as those working in critical care facilities; it is striking how the first cases of Nipah, SARS, MERS and influenza A subtype H5N1 were all first identified by clinical teams in critical care facilities.

An inclusive, whole-of-society approach is challenging, and the challenges may be magnified in a conflict or post-conflict zone. Wars and conflicts not only increase the risk of epidemics as people move to escape violence and health services become harder to maintain<sup>36</sup>, but also make public health responses vulnerable to interruption, thus making them less effective. Then, miscommunication, mistrust, disease and violence can fuel each other in a vicious cycle. Engaging local communities remains the highest priority, even in unstable contexts such as North Kivu and Ituri provinces of the Democratic Republic of the Congo (DRC)<sup>37</sup>, where an Ebola epidemic started in August 2018. It seems inevitable that responding to epidemics in politically unstable environments will become more common, and skilled negotiators and peacekeepers will have to be better integrated in response teams. Equally essential, therefore, will be an improved understanding of these challenging operational contexts among affected communities and external responders alike.

#### Social sciences

Social scientists have long applied their skills and knowledge in epidemic responses, although their roles have become more visible in recent years<sup>38</sup>. By focusing on communities, social science humanizes the epidemic response<sup>39</sup>, helps to increase understanding of context and may uncover associations between the context or local practices and the risk of transmission. The Social Science in Humanitarian Action Platform<sup>40</sup> has successfully produced rapid reports and

#### Box 2

# Precision public health

Precision medicine refers to the use of genomic sequencing to retrace the specific course of a disease in individual patients, with the aim of being able to choose the best treatment option for each person. In public health, the analogous idea of precisely directing the right intervention to the right population is equally appealing.

The potential of such an approach has been illustrated by the identification of two areas in the United States in 2016 that were at risk of Zika transmission89. Rather than the whole country, or even only Florida, being declared at risk, these two areas each measured less than 5 km<sup>2</sup>, and the response focused only on these specific neighbourhoods. By contrast, a campaign against yellow fever, also in 2016, defined risk 'at the level of entire nations'.

A broad interpretation of precision public health<sup>90</sup> incorporates many different types of data to increase the power of epidemiology<sup>91</sup>. Such data would not only include genomic information, but also satellite imaging, mobile phone data, social media use data and so on. For example, a study published in 2019 combined epidemiological surveillance data, travel surveys, parasite genetics and anonymized mobile phone data to measure the spread of malaria parasites in southeast Bangladesh<sup>92</sup>. A retrospective analysis of mobile phone call data in Sierra Leone from 2015 showed how it might have been used to assess the impact of travel restrictions on mobility during the Ebola epidemic<sup>46</sup>.

The principle of selecting the most relevant information from all available data seems within the scope of good epidemiological practice already. The challenge is recognizing and incorporating new types of data when they become available.

briefings on regions in which an epidemic has been identified, and the Global Research Collaboration for Infectious Disease Preparedness includes a social science research funders' forum to 'propel research in this area 41, acknowledging that its integration in the preparation and response to outbreaks is often missing or added as an afterthought to solve a problem that could have been forseen. There is still much to learn about how epidemic responders and social scientists can make the most of each other's expertise<sup>42</sup> and how data from social science can fit into the wider information architecture of epidemic response.

As an example, behavioural surveillance<sup>43</sup> will be critical in twenty-first century responses to disease outbreaks44. Just as behavioural surveillance to improve the understanding of HIV was crucial in identifying high-risk groups for HIV infection, so human behaviours will continue to be important as we respond to future infectious diseases. For instance, the Ebola virus outbreak in West Africa probably began before December 2013, but it took several months before hospital transmission and traditional burial practices were found to be the leading causes of its rapid spread.

#### **Emerging technologies**

The increasing prevalence of mobile phones, wireless internet connectivity and social media activity raises the possibility of using these tools to gather data for epidemiological studies, diagnostics<sup>45</sup>, population mobility during an Ebola epidemic<sup>46</sup> or influenza incidence in real time<sup>47</sup>. Future developments in predictive technology, machine learning and artificial intelligence will bring more opportunities to move towards 'precision public health' (Box 2).

The use of data from people is becoming strictly controlled, however, and it will be a challenge to persuade countries to invest in a new surveillance system, for example, before its general effectiveness has been demonstrated at a country level<sup>48</sup>. Even then, technology-based solutions should be integrated with community-based programmes and other existing epidemic preparedness and response systems because surveillance is more effective when standardized among different countries, districts and communities. To this end, suites of guidance and open-access standardized tools are being developed for reporting cases of disease, as well as consent forms, standard operating procedures and training materials<sup>49</sup>, properly validated diagnostic assays and access to quality-assurance panels in public<sup>50</sup> and veterinary<sup>51</sup> health. The rising trend of engaging citizens in data gathering is also welcome—the use of mosquito-recognition apps enables the collection of data far beyond the capacity of routine mosquito surveillance<sup>52</sup>. This way, citizens feed information into the public health system and the feedback loop offers a fast and direct way to provide citizens with details of potential actions that they can take.

As well as potentially supporting diagnosis and surveillance<sup>53</sup>, the fast-developing field of genomic epidemiology<sup>54</sup> can yield information to track the evolution of a virus such as Ebola during an epidemic<sup>55,56</sup>. There will be times when it can detect outbreaks better than traditional epidemiology, illustrating the need to have these tools available in the same toolbox. During the large Lassa fever outbreak in Nigeria in 2018, real-time genomic sequencing provided clear evidence that the rapid increase was not due to a single Lassa virus variant, nor attributable to sustained human-to-human transmission. Rather, the outbreak was characterized by vast viral diversity defined by geography, with major rivers acting as barriers to migration of the rodent reservoir<sup>57</sup>. These findings were crucial in containing the outbreak.

Developing and sustaining the capacity to conduct real-time sequencing with adequate bioinformatics analyses at regional and national levels will be challenging in low- and middle-income countries. Moreover, investments in relatively high-tech capacity (such as real-time sequencing) are competing with other, arguably more fundamental needs, such as equipment and training in primary laboratories. Political engagement must be nurtured between epidemics: it is not enough to offer technological and laboratory support during a crisis, even with the promise of building capacity, if the political will is not there. However, with proper preparation, and accessible and trusted data sharing and governance mechanisms, laboratories with limited resources may be able to leap-frog into the twenty-first century<sup>58,59</sup>.

#### **Research and development**

Vaccination is one of the most effective public health interventions and innovative strategies for research and development of vaccines. such as using ring vaccination as a trial design during Ebola epidemics since  $2015^{60-62}$ , must be encouraged. At the start of the 2013-2015 epidemic in West Africa, vaccine candidates were already in development, based on a long history of preclinical research, although a lot of work was still required to get clinical trials underway in time to be useful<sup>63</sup>. In 2015, when Zika was first internationally recognized as a pathogen that could cause birth defects<sup>64</sup>, there was hardly any research and no vaccines in late-stage development. Two-and-a-half years later, results from three phase I clinical trials had been reported 65, although challenges remained for further development. The lack of a profitable market for such products means that pharmaceutical companies lack the incentives to push this work between epidemics. Initiatives such as the Coalition for Epidemic Preparedness Innovations are attempting to positively disrupt financing models for vaccines against epidemic diseases<sup>66</sup>, and stockpiles of meningococcal vaccine, yellow fever vaccine and oral cholera vaccine are maintained by the International Coordinating Group to minimize potential delays due to limited manufacturing capacity<sup>67</sup>.

Similarly, if investigational treatments or vaccines are to be used as part of the response to an epidemic, ethical protocols<sup>68</sup> for managing informed consent and introducing them in clinical settings must be planned in advance with at-risk communities (Box 3). Trial designs<sup>69</sup>

#### Box 3

# **Epidemic ethics**

In 2016, the PREVENT project received Wellcome funding to provide ethics guidance "at the intersection of pregnancy, vaccines, and emerging and re-emerging epidemic threats"93. This was in response to the newly recognized association between infection with Zika virus during pregnancy and microcephaly in the newborn. Developing a vaccine was an obvious route to explore, but many researchers felt that they could not conduct clinical trials with pregnant women because it is generally assumed that the risk to the woman, the fetus or both outweighs any potential benefit. However, as Heyrana et al. argue: "Preventing pregnant women from participating in clinical trials is well intentioned but misguided."94.

PREVENT rapidly developed guidance for including pregnant women and their babies in Zika vaccine research<sup>95</sup>, and has since extended their scope to "a roadmap for the ethically responsible, socially just, and respectful inclusion of the interests of pregnant women in the development and deployment of vaccines against emerging pathogens."88.

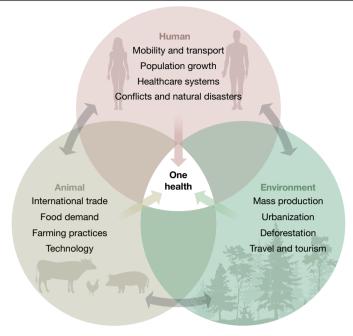
Integrating ethics in the preparation and response to epidemics does not close off avenues of research; it opens up possibilities and expedites progress.

should be created as soon as the option becomes viable. The essential consideration is how the resulting data can add to previous trials and influence the approach to trials in future epidemics. For example, research during the 2013-2015 Ebola epidemic enabled progress on therapeutic agents<sup>70</sup> that are now being trialled in the ongoing outbreak in DRC<sup>71</sup>. Scientific progress during and between epidemics must be matched by other workstreams, such as the preparation of supply chain logistics and communication with at-risk populations. Plans have to be made for a series of future outbreaks, enabling adaptive, multi-year, multi-country studies<sup>72</sup>. Similar plans are needed for continual preclinical research to ensure that future vaccine and therapeutic pipelines will be filled.

#### One Health

The term 'One Health'73 is used to acknowledge that human, animal and ecosystem health are tightly interconnected and need to be studied in the context of each other (Fig. 1). Changes in the environment—whether natural or anthropogenic—affect interactions between pathogens, vectors and hosts in multiple and complex ways, making the emergence or decline of endemic, epidemic and zoonotic diseases difficult to predict, while epidemics of animal diseases can challenge a community's access to food. The fact that pools of viruses, bacteria and parasites are maintained in wild and domesticated animals74 makes surveillance of potentially zoonotic diseases an intrinsic part of One Health epidemic planning. Many agencies and nations around the world now use prioritization tools such as those developed by the US CDC<sup>75</sup> or the United Nations (UN) Food and Agriculture Organization (FAO)<sup>76</sup> to identify and prioritize zoonotic diseases of concern. An early precedent was a joint consultation on emerging zoonotic diseases by the WHO, the FAO and the World Organisation for Animal Health in 2004<sup>77</sup>. Understanding disease ecology in the zoonotic reservoir could potentially lead to ways to predict the risk of human disease, thus providing the basis for smart early-warning surveillance systems.

Individual countries with limited resources for epidemiological studies and epidemic preparation and response must decide their own priorities. However, infectious diseases do not respect borders.



 $\textbf{Fig. 1} | \textbf{An ecosystem of interactions.} \ \textbf{The tightly interconnected nature of} \\$ human, animal and environmental health makes the emergence and decline of epidemics difficult to predict. One Health integrates multiple perspectives in a framework that emphasizes the need to consider any particular aspect in this broader context.

Similarly, the interdisciplinary nature of One Health means there are several different lenses through which different sectors assess risks and priorities. For One Health approaches to work, these multiple perspectives must be taken into account, whether human health or animal health, ecology or social sciences<sup>78</sup>.

#### Recovery

Epidemics do more than cause death and debilitation: they increase pressure on healthcare systems and healthcare workers and draw resources from services not directly linked to the epidemic. This can leave a legacy of distrust between people, governments and health systems, although more-positive outcomes have been found to strengthen relations between communities and public authorities. The full social and economic costs of the Ebola outbreak in West Africa have been estimated<sup>79</sup> to be as high as US\$53 billion when including the effect on health workers, long-term conditions suffered by 17,000 Ebola survivors, and costs of treatment, infection control, screening and deployment of personnel beyond West Africa. As healthcare resources became increasingly allocated to the Ebola response, hospital admissions fell and deaths from other diseases rose markedly, adding US\$18.8 billion to the estimated cost. Such pressure can be withstood in high-income countries with strong health systems, but in low-income countries the pressure can quickly reach a breaking point.

Ebola killed almost 1.5% of doctors, nurses and midwives in Guinea, 6.85% in Sierra Leone and just over 8% in Liberia 80. This is compared to mortality between 0.02% and 0.11% of the whole population of these countries. Estimates of the effect of this loss on maternal mortality suggest that thousands more women may have died in childbirth each year since the epidemic ended. Beyond the tragic deaths of so many healthcare workers, people were less likely to use health services for children or adults during the epidemic, suggesting decreased trust or even fear of healthcare settings<sup>81</sup>. More recently, in some areas affected by the 2018 Ebola outbreak in DRC, the introduction of free non-Ebola healthcare led to unprecedented demand. However, healthcare facilities

were not given sufficient additional resources to care for the number of people, which may have contributed to nosocomial infections.

Survivors, too, need to be cared for long after the epidemic is declared over. A cohort of more than 3,000 children is growing up in Brazil after being born with microcephaly because their mothers were infected with Zika during pregnancy. Tracking the development of these children increases understanding of the effects of Zika infection and helps to define what medical and social support the affected families may need as many of the children will grow up with severe developmental delays<sup>82</sup>.

#### **Outlook**

The challenges posed by twenty-first century epidemics are real and changing: future epidemics will be fuelled by conflict, poverty, climate change, urbanization and the broader demographic transition. In our response we must consider epidemics not as discrete events, but rather as connected cycles for which we can prepare, even if we cannot predict specific outbreaks. The challenge is then to choose the right response at the right scale in the right area at the right time. There needs to be a greater emphasis on absorbing and using positive lessons from each episode and avoiding those that led to negative outcomes<sup>83</sup>.

The way that we train practitioners and researchers working in all fields relevant to today's epidemic landscape has to change. A modern approach that is capable of characterizing epidemics and the best ways to control them must go beyond a narrow definition of epidemiology that sustains artificial barriers between disciplines. Instead, it must be able to integrate tools and practices from a diverse range of established and emerging scientific, humanistic, political, diplomatic and security fields. We believe that such an approach needs to become the norm for the curriculums of schools of public health around the world.

As well as training new generations of epidemiologists so that they have the skills, knowledge and networks to recognize and make use of every tool available to help them to do their work effectively, the entire architecture of the response to epidemics has to be adapted. Only then will we be able to maintain the comprehensive and effective response-including prevention and research-needed to stop epidemics and protect people's lives, no matter what the circumstances.

- Huxley, T. H. Nature: aphorisms by Goethe, Nature 1, 9-11 (1869).
- 2. Lilienfeld, D. E. Celebration: William Farr (1807-1883)—an appreciation on the 200th anniversary of his birth. Int. J. Epidemiol. 36, 985-987 (2007).
- Kadar, N. Rediscovering Ignaz Philipp Semmelweis (1818-1865). Am. J. Obstet. Gynecol. 220, 26-39 (2019).
- Julia, C. & Valleron, A.-J. Louis-Rene Villerme (1782-1863), a pioneer in social epidemiology: re-analysis of his data on comparative mortality in Paris in the early 19th century. J. Epidemiol. Community Health 65, 666-670 (2011).
- Fine, P. et al. John Snow's legacy: epidemiology without borders. Lancet 381, 1302-1311
  - This is a wide-ranging meeting report that places modern epidemiology in the context of the past two hundred years and highlights the importance of bringing in new disciplines, remaining open-minded and using those skills across a wider range of societal issues than are traditionally considered public health.
- Howard-Jones, N. Robert Koch and the cholera vibrio: a centenary, Br. Med. J. (Clin. Res. Ed.) 288, 379-381 (1984).
- Coggon, D., Rose, G. & Barker, D. J. P. Epidemiology for the Uninitiated (BMJ Books, 7. 2003).
- Omran, A. R. The epidemiologic transition. A theory of the epidemiology of population 8. change. Milbank Mem. Fund Q. 49, 509-538 (1971).
- Gallagher, J. Large Ebola outbreaks new normal, says WHO. BBC News (7 June 2019).
- Brownson, R. C., Samet, J. M. & Bensyl, D. M. Applied epidemiology and public health: are we training the future generations appropriately? Ann. Epidemiol. 27, 77-82 (2017).
- 11 WHO. Managing Epidemics: Key Facts about Major Deadly Diseases https://apps.who.int/ iris/handle/10665/272442 (WHO, 2018).
- 12. Carter, K. C. Ignaz Semmelweis, Carl Mayrhofer, and the rise of germ theory. Med. Hist. 29, 33-53 (1985).
- Plotkin, S. History of vaccination. Proc. Natl Acad. Sci. USA 111, 12283-12287 (2014). 13
- Strassburg, M. A. The global eradication of smallpox. Am. J. Infect. Control 10, 53-59
- Nathanson, N. & Kew, O. M. From emergence to eradication: the epidemiology of poliomyelitis deconstructed. Am. J. Epidemiol. 172, 1213-1229 (2010).
- Macfarlane Burnet, F. & White, D. O. Natural History of Infectious Disease p263 (Cambridge Univ. Press, 1972).

- Bregman, D. J. & Langmuir, A. D. Farr's law applied to AIDS projections. J. Am. Med. Assoc. 263, 1522-1525 (1990)
- Smith, G. D. & Ebrahim, S. Epidemiology—is it time to call it a day? Int. J. Epidemiol. 30, 1-11 (2001).
- Smith, K. F. et al. Global rise in human infectious disease outbreaks. J. R. Soc. Interface 11, 20140950 (2014).
- MacIntyre, C. R. & Bui, C. M. Pandemics, public health emergencies and antimicrobial resistance — putting the threat in an epidemiologic and risk analysis context, Arch. Public Health 75, 54 (2017).
- Neiderud, C.-J. How urbanization affects the epidemiology of emerging infectious diseases. Infect. Ecol. Epidemiol. 5, 27060 (2015).
- Morse, S. S. in Microbial Evolution and Co-Adaptation: A Tribute to the Life and Scientific Legacies of Joshua Lederberg (National Academies Press, 2009).
- Vignier, N. & Bouchaud, O. Travel, migration and emerging infectious diseases. EJIFCC 29, 175-179 (2018).
- Paaijmans, K. P., Read, A. F. & Thomas, M. B. Understanding the link between malaria risk and climate. Proc. Natl Acad. Sci. USA 106, 13844-13849 (2009).
- Boozary, A. S., Farmer, P. E. & Jha, A. K. The Ebola outbreak, fragile health systems, and quality as a cure. J. Am. Med. Assoc. 312, 1859-1860 (2014).
- Quinn, S. C. & Kumar, S. Health inequalities and infectious disease epidemics: a challenge for global health security. Biosecur, Bioterror, 12, 263-273 (2014).
- Cohn. S. & Kutalek, R. Historical parallels. Ebola virus disease and cholera: understanding community distrust and social violence with epidemics. PLoS Curr. 8, https://doi.org/10.1371/ currents.outbreaks.aa1f2b60e8d43939b43fbd93e1a63a94 (2016)
- Sharara, S. L. & Kanj, S. S. War and infectious diseases: challenges of the Syrian civil war. PLoS Pathog. 10, e1004438 (2014).
- WHO. Pandemic Influenza Preparedness Framework for the Sharing of Influenza Viruses and Access to Vaccines and Other Benefits (WHO, 2011).
- 30. Nkengasong, J. N. How Africa can quell the next disease outbreaks. Nature 567, 147 (2019).

The ability to prevent, detect and respond to any health issues will always depend on the local capacity and although international partners can bring complementary expertise and resources, it is the local capacity that is critical; in this article, the authors argue for national investment in public health, health systems, science and local leadership, examples of which are the establishment of the African CDC, the renewed strength of the WHO African regional office and the African Academy of Sciences.

- Cheng, M. AP Exclusive: UN health chief orders probe into misconduct. AP News (17 January 2019).
- Tudor Hart, J. The inverse care law. Lancet 297, 405-412 (1971).
- Kay, S. Africa's leadership in biomedical research: shifting the center of gravity. Sci. Transl. Med. 7, 314ed13 (2015).

Agenda setting, research questions and funding for biomedical research has historically been led from Northern Hemisphere countries in an unequal Northern-Southern Hemisphere relationship: in this article, a determined approach to shift that centre of gravity is outlined, such that the agenda is firmly based where the need is greatest

- Chataway, J. et al. Science granting councils in sub-Saharan Africa: trends and tensions. Sci. Public Policy 46, 620-631 (2019).
- 35. International Federation of Red Cross and Red Crescent Societies. Community-Based Surveillance: Guiding Principles (IFRC, 2017).
- 36. Gayer, M., Legros, D., Formenty, P. & Connolly, M. A. Conflict and emerging infectious diseases. Emerg. Infect. Dis. 13, 1625-1631 (2007).
- Vinck, P., Pham, P. N., Bindu, K. K., Bedford, J. & Nilles, E. J. Institutional trust and misinformation in the response to the 2018-19 Ebola outbreak in North Kivu, DR Congo: a population-based survey. Lancet Infect. Dis. 19, 529-536 (2019).
- Bedford, J. et al. Application of social science in the response to Ebola, Équateur Province, Democratic Republic of the Congo. Wkly. Epidemiol. Rec. 94, 19-23
- Bardosh, K. et al. Towards People-Centred Epidemic Preparedness and Response: From Knowledge to Action (Wellcome Trust, 2019).
- UNICEF & IDS. The Social Science in Humanitarian Action Platform https://www. socialscienceinaction.org/ (2019).
- GloPID-R. Social Science Research. https://www.glopid-r.org/our-work/social-scienceresearch/ (2018).
- 42. Stellmach, D., Beshar, L., Bedford, J., du Cros, P. & Stringer, B. Anthropology in public health emergencies; what is anthropology good for? BMJ Glob, Health 3, e000534
- Manhart, L. E. & Khosropour, C. M. Launching a new era for behavioural surveillance. Sex. Transm. Infect. 91, 152-153 (2015).
- Miller, M. & Hagan, E. Integrated biological-behavioural surveillance in pandemic-threat warning systems. Bull. World Health Organ. 95, 62-68 (2017).
- Wood, C. S. et al. Taking connected mobile-health diagnostics of infectious diseases to the field, Nature 566, 467-474 (2019).
- Peak, C. M. et al. Population mobility reductions associated with travel restrictions during the Ebola epidemic in Sierra Leone: use of mobile phone data. Int. J. Epidemiol. 47, 1562-1570 (2018).
- Broniatowski, D. A., Paul, M. J. & Dredze, M. National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. PLoS ONE 8, e83672
- Velasco, E., Agheneza, T., Denecke, K., Kirchner, G. & Eckmanns, T. Social media and internet-based data in global systems for public health surveillance: a systematic review. Milbank Q. 92, 7-33 (2014).
- ISARIC. Protocols & Data Tools https://isaric.tghn.org/protocols/ (accessed 18 September 2019).
- WHO. Laboratory Network. https://www.who.int/immunization/monitoring\_surveillance/ burden/laboratory/en/ (2018)
- FAO. Enhancing Early Warning Capabilities and Capacities for Food Safety (FAO, 2015).

#### Review

- European Citizen Science Association. Global Mosquito Alert https://ecsa.citizen-science. net/global-mosquito-alert (accessed 12 October 2019)
- 53. Quick, J. et al. Real-time, portable genome sequencing for Ebola surveillance. Nature **530**, 228-232 (2016).
- Tang, P., Croxen, M. A., Hasan, M. R., Hsiao, W. W. & Hoang, L. M. Infection control in the new age of genomic epidemiology. Am. J. Infect. Control 45, 170-179 (2017).
- Gire, S. K. et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science 345, 1369-1372 (2014).
- 56. Tong, Y.-G. et al. Genetic diversity and evolutionary dynamics of Ebola virus in Sierra Leone, Nature **524**, 93–96 (2015).
- Siddle, K. J. et al. Genomic analysis of Lassa virus during an increase in cases in Nigeria in 2018. N. Engl. J. Med. **379**, 1745-1753 (2018).
- Aarestrup, F.M. et al. Integrating genome-based informatics to modernize global disease monitoring, information sharing, and response, Emerg. Infect. Dis. 18, e1 (2012) The integration of genomics and other types of data into the surveillance, prevention and response of epidemics is critical and can help to transform the ability to enhance public health: although the tools are now available, it will be key to ensure that these new approaches are fully integrated and not seen as esoteric ivory tower research, but instead as an essential component of twenty-first century epidemiology, public health and epidemics—the next generation of leaders need to be trained in and comfortable
- GMI Global Microbial Identifier https://www.globalmicrobialidentifier.org/ (2018).

across a range of new disciplines.

- Henao-Restrepo, A. M. et al. Efficacy and effectiveness of an rVSV-vectored vaccine in preventing Ebola virus disease: final results from the Guinea ring vaccination, open-label, cluster-randomised trial (Ebola Ça Suffit!). Lancet 389, 505-518 (2017).
  - A seminal study that shows that ring vaccination could be used in the midst of a devastating Ebola epidemic and, furthermore, that innovation research can be conducted in an epidemic, trial designs can be adapted without compromising scientific integrity and that Ebola can be prevented through vaccination.
- Rid, A. & Miller, F. G. Ethical rationale for the Ebola "ring vaccination" trial design. Am. J. Public Health 106, 432-435 (2016).
- Wells, C. R. et al. Ebola vaccination in the Democratic Republic of the Congo. Proc. Natl Acad. Sci. USA 116, 10178-10183 (2019).
- Rojek, A., Horby, P. & Dunning, J. Insights from clinical research completed during the West Africa Ebola virus disease epidemic. Lancet Infect. Dis. 17, e280-e292 (2017).
- Siedner, M. J., Ryan, E. T. & Bogoch, I. I. Gone or forgotten? The rise and fall of Zika virus. Lancet Public Health 3, e109-e110 (2018).
- 65. Barrett, A. D. T. Current status of Zika vaccine development: Zika vaccines advance into clinical evaluation. NPJ Vaccines 3, 24 (2018).
- 66 Burki, T. CEPI: preparing for the worst, Lancet Infect, Dis. 17, 265-266 (2017)
- 67. Yen, C. et al. The development of global vaccine stockpiles. Lancet Infect. Dis. 15, 340-347 (2015).
- 68. Calain, P. The Ebola clinical trials: a precedent for research ethics in disasters, J. Med. Fthics 44 3-8 (2018)
  - It is an ethical imperative to consider and implement research in an epidemic setting as, for many epidemic diseases, it is the only time at which to conduct the research that will inform and improve the lives of the individuals affected during epidemic and to ensure that future generations are better prepared; however, such research is challenging at many levels and it is critical to have an ethical framework that guides the research, places individuals and communites at the heart of the research and facilitates the maximum benefit for the maximum number of people, in an equitable way, that is independent of the ability to pay-such a framework is outlined in this paper and is put into the context of social justice and equity
- 69. Lipsitch, M. & Eyal, N. Improving vaccine trials in infectious disease emergencies. Science **357**, 153-156 (2017)
- Mendoza, E. J., Qiu, X. & Kobinger, G. P. Progression of Ebola therapeutics during the 2014-2015 outbreak. Trends Mol. Med. 22, 164-173 (2016).
- Fallah, M. P. & Skrip, L. A. Ebola therapies: an unconventionally calculated risk. Lancet 393, 850-852 (2019).
- Brueckner, M., Titman, A., Jaki, T., Rojek, A. & Horby, P. Performance of different clinical trial designs to evaluate treatments during an epidemic. PLoS ONE 13, e0203387 (2018).
- 73. Destoumieux-Garzón, D. et al. The One Health concept: 10 years old and a long road ahead, Front, Vet. Sci. 5, 14 (2018).
- Day, M. J. et al. Surveillance of zoonotic infectious disease transmitted by small companion animals, Emerg. Infect. Dis. 18, https://doi.org/10.3201/eid1812.120664 (2012).
- Rist, C. L., Arriola, C. S. & Rubin, C. Prioritizing zoonoses: a proposed one health tool for collaborative decision-making. PLoS ONE 9, e109986 (2014).
- 76. FAO. Evaluation of the Emergency Prevention System (EMPRES) Programme in Food Chain Crises (FAO, 2018).

- FAO, WHO & OIE, Report of the WHO/FAO/OIE ioint consultation on emerging zoonotic diseases (WHO, 2004)
- 78. European Centre for Disease Prevention and Control. Towards One Health preparedness (ECDC, 2018).
- Huber, C., Finelli, L. & Stevens, W. The economic and social burden of the 2014 Ebola outbreak in West Africa. J. Infect. Dis. 218, S698-S704 (2018). Epidemics cause enormous disruption to countries, regions and the world; however, the focus is often on the epidemic itself, the pathogen and its immediate effect rather than the much broader effect that the epidemic has not only on the healthcare system—which lasts long after the epidemic itself—as routine vaccination programmes often collapse, maternal-child health suffers, and malaria, HIV and tuberculosis clinics and surgery—all aspects of healthcare—are disrupted, but also on the wider society, as mistrust and tension occurs between citizens, authorities and
- already fragile communities. Evans, D. K., Goldstein, M. & Popova, A. Health-care worker mortality and the legacy of the Ebola epidemic, Lancet Glob, Health 3, e439-e440 (2015).

governments, and education, investments, businesses, trade and tourism inevitablely

suffer leading to an economic impact that can be long lasting and devastating for often

- Morse, B., Grépin, K. A., Blair, R. A. & Tsai, L. Patterns of demand for non-Ebola health services during and after the Ebola outbreak: panel survey evidence from Monrovia, Liberia. BMJ Glob. Health 1, e000007 (2016).
- Brickley, E. B. & Rodrigues, L. C. Further pieces of evidence in the Zika virus and microcephaly puzzle. Lancet Child Adolesc. Health 2, 162-164 (2018).
- 83. Ebola Gbalo Research Group. Responding to the Ebola virus disease outbreak in DR Congo: when will we learn from Sierra Leone? Lancet 393, 2647-2650 (2019).
- Miller, D. D. & Brown, E. W. Artificial intelligence in medical practice: the question to the answer? Am. J. Med. 131, 129-133 (2018).
- Benke, K. & Benke, G. Artificial intelligence and big data in public health. Int. J. Environ. Res. Public Health 15, 2796 (2018).
- McLellan, J. S. et al. Structure of RSV fusion glycoprotein trimer bound to a prefusionspecific neutralizing antibody. Science 340, 1113-1117 (2013).
- Adalja, A. A., Watson, M., Cicero, A. & Inglesby, T. Vaccine Platforms: State of the Field and Looming Challenges (Johns Hopkins Center for Health Security, 2019).
- Charlton Hume, H. K. & Lua, L. H. L. Platform technologies for modern vaccine manufacturing. Vaccine 35, 4480-4485 (2017).
- Dowell, S. F., Blazes, D. & Desmond-Hellmann, S. Four steps to precision public health. Nature 540, 189-191 (2016).
- 90. Chowkwanyun, M., Bayer, R. & Galea, S. "Precision" public health — between novelty and hype. N. Engl. J. Med. 379, 1398-1400 (2018).
  - Horton, R. Offline: in defence of precision public health. Lancet 392, 1504 (2018).
- Chang, H. H. et al. Mapping imported malaria in Bangladesh using parasite genetic and 92. human mobility data, eLife 8, e43481 (2019).
- Krubiner, C. B. et al. Pregnant women & vaccines against emerging epidemic threats: ethics guidance for preparedness, research, and response. Vaccine https://doi.org/ 10.1016/i.vaccine.2019.01.011 (2019).
- Heyrana, K., Byers, H. M. & Stratton, P. Increasing the participation of pregnant women in clinical trials. J. Am. Med. Assoc. 320, 2077-2078 (2018).
- The Ethics Working Group on ZIKV Research & Pregnancy. Pregnant Women & the Zika Virus Vaccine Research Agenda: Ethics Guidance on Priorities, Inclusion, and Evidence Generation (PREVENT, 2017).

Acknowledgements We thank M. Regnier at Wellcome for editing the manuscript.

Author contributions All authors developed the scope and focus of the Review and contributed to the writing of the manuscript.

Competing interests The authors declare no competing interests.

#### Additional information

Correspondence and requests for materials should be addressed to J.F. Reviewer information Nature thanks Peter Byass, Sharon Peacock and the other, anonymous, reviewer(s) for their contribution to the peer review of this work

Reprints and permissions information is available at http://www.nature.com/reprints. Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© Springer Nature Limited 2019

# Sex and gender analysis improves science and engineering

https://doi.org/10.1038/s41586-019-1657-6

Received: 8 January 2019

Accepted: 27 August 2019

Published online: 6 November 2019

Cara Tannenbaum<sup>1,7</sup>, Robert P. Ellis<sup>2,7</sup>, Friederike Eyssel<sup>3,7</sup>, James Zou<sup>4,5,7</sup> & Londa Schiebinger6\*

The goal of sex and gender analysis is to promote rigorous, reproducible and responsible science. Incorporating sex and gender analysis into experimental design has enabled advancements across many disciplines, such as improved treatment of heart disease and insights into the societal impact of algorithmic bias. Here we discuss the potential for sex and gender analysis to foster scientific discovery, improve experimental efficiency and enable social equality. We provide a roadmap for sex and gender analysis across scientific disciplines and call on researchers, funding agencies, peer-reviewed journals and universities to coordinate efforts to implement robust methods of sex and gender analysis.



Integrating sex and gender analysis into the design of research, where relevant, can lead to discovery and improved research methodology. A deeper understanding of the genetic and hormone-mediated basis for sex differences in immunity, for example, promises insights into novel cancer immunotherapies<sup>1</sup>. Evidence that facial recognition systems misclassify gender more often for darker-skinned women than for lighter-skinned men has led to refinements in computer vision<sup>2</sup>. Understanding sex-based responses to climate change allows better modelling of demographic change among marine organisms and the downstream effects for humans<sup>3,4</sup>. Sex or gender analysis can be critical to the interpretation, validation, reproducibility and generalizability of research findings (Box 1).

The documented importance of sex and gender analysis in research has underwritten policy change at major funding agencies. New policies have been implemented at the Canadian Institutes of Health Research (2010), European Commission (2014), US National Institutes of Health (2016), German Research Foundation (2020), among others. Concurrently, peer-review journals have implemented editorial guidelines to evaluate the rigour of sex and gender analysis as one criterion among many when selecting manuscripts for publication. The goal is to increase transparency, promote inclusion and reset the research default to carefully consider sex and gender, where appropriate.

In this Perspective, we discuss how incorporating sex and/or gender analysis into research can improve reproducibility and experimental efficiency, help to reduce bias, enable social equality in scientific outcomes and foster opportunities for discovery and innovation. From highlighted examples, we extract decision-tree roadmaps for researchers across disciplines. We consider the limits to sex and gender analysis and offer recommendations to researchers and funding agencies on how to move the field forward. Throughout this Perspective, we explore how integrating sex and gender analysis into research design has the potential to offer new perspectives, pose new questions and, importantly, enhance social equalities by ensuring that research findings are applicable across the whole of society.

#### Reproducibility and efficiency

Reproducibility is important for scientific excellence. One important reason for a lack of reproducibility in experimentation is inconsistency in methodological reporting, which varies widely across disciplines from biology to chemistry, human-robot interaction, medicine, physics, psychology and beyond<sup>5,6</sup>. Sex- and gender-specific reporting is still limited in a range of scientific disciplines. In preclinical microbiology and immunology, a review of published studies using primary cells from diverse animal species (that is, humans and nonhuman vertebrates) revealed that the majority failed to report the sex of donors from which the cells were isolated<sup>7,8</sup>. In marine science, a review of experimental ocean acidification studies showed that only 3.9% of studies statistically assessed sex-based differences, while only 10.5% of studies accounted for possible sex effects by assessing females and males independently<sup>9</sup>. Similarly, in ecotoxicology, a review of omics studies showed that although most reported sex, only 23% (5 out of 22) examined the omics response of each sex to a toxicant<sup>10</sup>. In social robotics, the notion of robot gender, genderstereotypical domains and their interaction with user gender has only recently become a target of scientific inquiry<sup>11</sup>. A lack of transparency in reporting sex and gender-related variables makes it difficult to reproduce experiments in which these variables affect experimental results.

#### Disaggregating the data

Analysing experimental results by sex and/or gender is critical for improving accuracy and avoiding misinterpretation of data (Fig. 1). The common practice of pooling the response of females and males or

<sup>1</sup>Institute of Gender and Health, Canadian Institutes of Health Research, Université de Montréal, Montreal, Quebec, Canada. <sup>2</sup>College of Life and Environmental Sciences, University of Exeter, Exeter, UK, 3 Center of Excellence Cognitive Interaction Technology, Department of Psychology, Universität Bielefeld, Bielefeld, Germany, 4 Biomedical Data Science, Stanford University, Stanford, CA, USA. 5Chan-Zuckerberg Biohub, San Francisco, CA, USA. 6History of Science, Gendered Innovations in Science, Health & Medicine, Engineering and Environment, Stanford University, Stanford, CA, USA, 7These authors contributed equally: Cara Tannenbaum, Robert P. Ellis, Friederike Eyssel, James Zou, \*e-mail; schiebinger@stanford.edu

## **Perspective**

#### Box 1

# Distinguishing sex and gender

Sex refers to the biological attributes that distinguish organisms as male, female, intersex (ranging from 1:100 to 1:4,500 in humans, depending on the criteria used<sup>126,127</sup>) and hermaphrodite (over 30% of noninsect nonhuman animals<sup>128</sup>). In biology, sex describes differences in sexual characteristics within plants or animals that go beyond their reproductive functions to affect appearance, physiology or neuroendocrine, behavioural and metabolic systems. In engineering, sex includes anthropometric, biomechanical and physiological characteristics that may affect the design of products, systems and processes.

Gender refers to psychological, social and cultural factors that shape attitudes, behaviours, stereotypes, technologies and knowledge. Gender includes three related dimensions. Gender norms refer to spoken and unspoken rules in the family, workplace, institution or global culture that influence individuals. Gender identity refers to how individuals and groups perceive and present themselves within specific cultures. Gender relations refer to power relations between individuals with different gender roles and identities<sup>129</sup>.

Sex and gender interact in unexpected ways. Pain, for example, exhibits biological sex differences in the physiology of signalling. Pain also incorporates sociocultural components in how symptoms are reported by women, men and gender-diverse people, and how physicians understand and treat pain according to a patient's gender<sup>130</sup>.

women and men can mask sex differences. For example, consider copepods, small aquatic crustaceans. Failure to disaggregate and analyse data by sex leads to the false interpretation that increased levels of  $p_{\rm CO_2}$  have no significant biological effect on respiration (Fig. 1b). By contrast, disaggregating data by sex reveals important sex-based differences in the respiration rate of females and males in response to increased  $p_{\rm CO_2}$  levels  $^{12}$ .

The same is true for human research. Pooling data yields inexact results. In a human–robot experiment, humans were asked to touch or point to anatomical regions on a 59-cm NAO robot. When asked to touch accessible regions (such as hands and feet), there was little physiological reaction; when asked to touch inaccessible regions (such as the plastic buttocks or genitals of the robot), human participants had increased heart rate and blood pressure<sup>13</sup>. Equal numbers of women and men were recruited for the experiment; however, the data were not disaggregated or analysed separately. We know that norms for human social touch vary according to the age, gender identity and cultural background of the participant—as well as social context and purpose of the touch 14. If results are not stratified by these variables, opportunities will be missed to provide clearer insights into their influence on human judgments and behaviour.

#### Variability, sample size and interactions

Scientists have erroneously assumed that females should be excluded from experiments because of the variable nature of the data caused by the reproductive cycle<sup>15</sup>. In fact, research has shown that males exhibit equal or greater variability than females for specific traits owing to fluctations in testosterone levels and other factors, such as animal group caging<sup>16</sup>. Analysis of microarray datasets reveals similar findings that females are no more variable than males on measures of gene expression in both mice and humans<sup>17</sup>. Accounting for sex and gender enhances the

likelihood of detecting meaningful effects, elucidating unexplained variability and potentially reducing the overall number of experiments required to determine trends or make ground-breaking discoveries. In a meta-analysis of 11 proteomics datasets from humans and mice, sex explained 13.5% of the observed variation of complex protein abundances and stoichiometry, even more than other environmental factors, such as diet<sup>18</sup>.

On the surface, it may appear that including females and males, women and men in a study necessitates doubling the number of experimental participants. However, this is not always the case. More efficient experimental designs can incorporate both sex and gender while maintaining control over variance  $^{19}$ . Factorial designs, in which two experimental factors with multiple levels are tested, and data are collected across all possible combinations of factors and levels, are one such strategy. This enables the effect of each factor to be tested, in addition to the interaction between the factor levels. For such cases, sample sizes may need to be slightly increased by 14-33% to account for the extra parameter being estimated, but they do not need to be doubled, according to sample size calculators that consider interaction effects  $^{20,21}$ . Analysing data by sex or gender enhances the likelihood of detecting meaningful effects that, in turn, help to reduce confounding, increase reproducibility and reduce the cumulative number of experiments required.

Numerous interactions, such as the interaction of the sex of the research participants, may also influence outcomes. In animal research, females and males are often studied separately in the laboratory. Yet in the wild, the sexes coexist—and their interactions can influence research results. Recent studies of longevity in the nematode, *Caenorhabditis elegans*, found that the presence of males accelerated ageing in individuals of the opposite sex (in this case, hermaphrodites). In other words, hermaphrodites died at a younger age in the presence of males. Researchers traced this 'male-induced demise' to pheromones released by males and found it could occur without mating and required only that the hermaphrodites be exposed to the medium in which males were once present<sup>22</sup>. Ignoring such interactions potentially leads to an incomplete understanding of species viability in the wild.

Other interactions focus on the sex of the researcher and potential impacts on research participants. In social science, it has long been understood that the simple presence of an observer can alter the response of the observed, whether in the field or in laboratory experiments<sup>23</sup>. In quantum mechanics, the act of observation can alter the phenomenon by collapsing the wave function. Similarly, in animal research, experimenter sex can influence research outcomes. A study exploring pain showed that rats and mice did not exhibit pain when a male experimenter was present, as opposed to when a female experimenter was present in the room or when in an empty room. Both female and male mice displayed this 'male observer' effect, but female mice did so to a greater extent. Researchers determined that the mice responded to male-associated olfactory stimuli<sup>24</sup>. The authors suggest that not controlling for experimenter sex throws into question many of the previously published studies on pain research.

Many other examples of these types of interactions—crucial to excellence and discovery in research—could be discussed. However, here we would like to include one further interaction of note, namely of researcher gender and the type of research conducted. Two studies provide compelling evidence that in biomedical, clinical and public health research, women in leading positions (first and last author) are more likely to analyse sex and gender in published research <sup>25,26</sup>. However, this dynamic has not yet been replicated in other research fields, such as computer science, engineering or the physical sciences.

#### **Opportunities for discovery**

Ignoring sex and gender analysis can lead to inaccuracies, research inefficiency and difficulties generalizing results. Integrating sex and gender analysis into research can open the door to discovery and innovation.

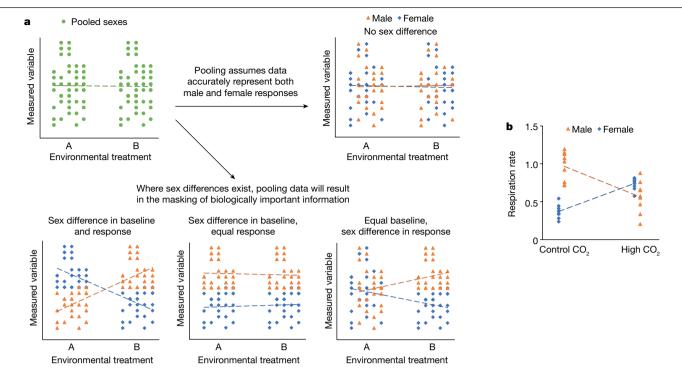


Fig. 1 | Hazards of pooling data from both sexes. Pooling data across sexes not only assumes that there is no difference between males and females, but also subsequently prevents researchers from testing for the dependency of an experimental response on the sex of a study participant. a, The theoretical examples reveal that pooling (green circles) masks important male (orange triangles) and female (blue squares) differences in baseline data, treatment

response and sex × treatment interactions—any one of which leads to  $misinterpretation of the results. \, \boldsymbol{b}, An \, example \, of \, experimental \, data \, in \, which$ pooling would have masked both the sex difference in the respiration rate of copepods, as well as the response of this variable to increased levels of  $p_{CO_2}$ . Theoretical examples were generated using hypothetical data; experimental data were taken from a previously published study<sup>12</sup>.

A prevalent assumption is that sex is a binary trait determined genetically before birth, and that it is fixed across lifespan<sup>27,28</sup>. Commonly used model organisms in biology, such as mice, Drosophila melanogaster and C. elegans, reinforce these perceptions. Sex, however, can be highly plastic, and studying interactions with the environment, for example, has led to new understandings of the mechanisms of sex determination within the context of global climate change.

The sex ratio of a population influences its resilience to environmental disturbances. The mechanism that determines sex is thus a vital consideration for predicting population viability<sup>29,30</sup>. Enhancing the capacity of sex analysis for a growing number of species, across a wide range of settings, may increase our ability to accurately model the effects of climate change.

#### Climate impacts in the ocean

For species reliant on temperature for sex determination, rapid global warming poses a risk to sex ratios and demographic stability. Turtles are the most widely studied group in which sex is determined by temperature. The ability to differentiate between female and male juvenile green sea turtles using non-invasive endocrine markers has enabled the discovery that global warming negatively skews population sex ratios. Turtles originating from warmer northern Great Barrier Reef sites, for instance, exhibit a female sex ratio of 99%, whereas cooler southern sites maintain a 68% female juvenile ratio<sup>3</sup>. Similarly, in fish species with temperature-dependent sex determination, warming is projected to result in male-skewed populations (up to 3:1 male:female) by the end of the century28. Such changes in sex balance can limit mate choice, reduce reproductive capacity and undermine population viability<sup>31,32</sup>.

Warming is not occurring in isolation, but against a backdrop of anthropogenic disturbances across marine environments, which include habitat destruction, pollution and overfishing. Primary sex differentiation has been shown to respond to a diverse range of these environmental factors in a growing number of species. Hypoxia, for example, has resulted in a higher ratio of males in zebrafish<sup>33</sup>. Similarly, ocean acidification results in 16% more female oysters over a single generational cycle<sup>4</sup>, and increased aquatic pH results in more female cichlids<sup>34</sup>. What is increasingly apparent is that alterations in sex ratio in either direction—will result in populations that are less resilient to further disturbance and potentially lead to demographic collapse<sup>35,36</sup>.

Social organization can also influence population sex ratios. Numerous nonhuman species develop elaborate social organizations, and sex determination can be socially mediated. Clownfish, for example, are protandrous hermaphrodites (they mature as male; some change to female) that live in a strict social hierarchy with a single dominant and highly fecund female at the top who mates with a single large male in the social group; all remaining individuals remain immature juveniles. Removal of the alpha female results in the alpha male changing sex to female, with all subordinates moving up a rung in the social hierarchy<sup>37</sup>. By contrast, many grouper species, a subfamily of long-lived and high-value reef species, are protogynous hermaphrodites (they mature as female; some change to male). Large dominant males control groups of females with strong sexual selection, resulting in these males achieving the greatest reproductive success. These sequentially hermaphroditic individuals consistently produce more offspring and enjoy greater reproductive success after they have changed sex<sup>36</sup>. Thus, the timing and the direction of sex change are crucial species $specific factors \, that \, determine \, demographic \, resilience \, to \, disturbance \, in \, determine \, demographic \, resilience \, to \, disturbance \, in \, determine \, demographic \, resilience \, to \, disturbance \, in \, determine \, demographic \, resilience \, to \, disturbance \, in \, determine \, demographic \, resilience \, to \, disturbance \, in \, determine \, demographic \, resilience \, to \, disturbance \, in \, determine \, demographic \, resilience \, to \, disturbance \, in \, determine \, demographic \, resilience \, to \, disturbance \, in \, determine \, demographic \, resilience \, to \, disturbance \, in \, determine \, demographic \, resilience \, demographic \, resilience \, to \, disturbance \, in \, determine \, demographic \, resilience \, res$ sex-changing organisms.

A mechanistic understanding of these and other ecologically important sex-based responses enables more accurate modelling of the effects of environmental variability, climate change or anthropogenic disturbance (for example, overfishing) at a population level. Sex-specific

## **Perspective**

effects of climate change stressors on sex determination mechanisms, particularly in commercially important species, have potentially important implications for humans with respect to aquatic food production, ecosystem services and biodiversity. Incorporating sex analysis into marine science—and the natural sciences more widely—enhances research excellence and opportunities for discovery.

#### **Targeted human therapeutics**

Sex analysis also reveals opportunities for human drug development. In the areas of pain and depression, the discovery of sex differences in molecular pathways has signalled new directions for targeted therapies<sup>38</sup>. Pain research that uses experimental mouse models of chronic pain shows that male and female mice withdraw from painful stimuli in a similar fashion, except when the contribution of microglial cells is inhibited<sup>39</sup>. Microglia are specialized immune cells located exclusively in the spinal cord and the brain. Inhibitors of microglia reduce pain sensing in male—but not female—mice, underscoring the potential importance of sex-dependent molecular pain pathways. Mouse models of depression also show sexually divergent networks in the brain with distinct patterns of stress-induced gene regulation in males and females<sup>40</sup>. These findings have now been reproduced in human postmortem tissue and may provide insights into why males and females with major depressive disorder respond differently to treatment with antidepressants<sup>40</sup>.

Although sex-specific dosages are rare, a few already exist. Such is the case for the drug desmopressin that activates vasopressin receptors in the kidney to regulate water homeostasis. Because the gene for the arginine vasopressin receptor is found on the X chromosome in a region that is likely to escape X-chromosome inactivation, women are more sensitive to the antidiuretic effects of vasopressin than men, who have only one X chromosome and therefore only one copy of the vasopressin receptor gene per cell<sup>41</sup>. As a result, older women who take desmopressin are more likely to experience a reduced sodium concentration in the blood than men, which corresponds to a higher incidence of side effects in women. To avoid unnecessary harm, both the European Union and Canada have recommended lower dosages for older women taking desmopressin.

Even cancer immunotherapy is benefitting from a deeper understanding of previously recognized genetic and hormone-mediated sex differences in immunity. Patients with melanoma or lung cancer, who are treated with checkpoint inhibitors, respond differently based on their sex, with a higher proportion of male than female patients achieving successful remission¹. Designed to outsmart the defence tactics of the cancer cells, checkpoint inhibitors stimulate natural killer cells to attack tumour cells. Natural killer cells are sensitive to oestrogen and testosterone, which may explain these observed sex differences. Understanding the underlying mechanisms will enable us to fine-tune future therapies⁴².

We expect to see an exponential rise in biomedical discoveries now that new computational biology and statistical genetics software facilitates the exploration of X-chromosome-related expression in complex diseases  $^{43}$ . Until recently, sex chromosomes were excluded from most genome-wide association studies because of the difficulty in distinguishing the active from the inactive X chromosome in females, and because of a mismatch in chromosomal size  $^{44,45}$ —the X chromosome has 1,669 known genes and the smaller Y chromosome contains only 426. Including sex chromosomes in genome-wide association studies, as well as including and analysing adequate numbers of female and male cells, tissues, animals and humans in research, will broaden our understanding of why women and men are affected differently by certain diseases and how we can adapt life-saving therapies to their specific needs.

#### **Engineering for equality**

An often neglected but crucial component of engineering is to understand the broader social impacts of the technology being developed and

to ensure that the technology enhances social equality by benefitting diverse populations. Human bias and stereotypes can be perpetuated, and even amplified, when researchers fail to consider how human preferences and assumptions may consciously or unconsciously be built into science or technology. Gender norms, ethnicity and other biological and social factors shape and are shaped by science and technology in a robust cultural feedback loop  $^{46}$ . This section discusses examples from product design, artificial intelligence (AI) and social robotics to illustrate how sex and gender analysis can enhance excellence in engineering.

#### **Designing safer products**

When products are designed based on the male norm, there is a risk that women and people of smaller stature will be harmed. Motor vehicle safety systems provide one such example. Because male drivers have historically been overrepresented in traffic data, seatbelts and airbags have been designed and evaluated with a focus on the typical male occupant with respect to anthropometric size, injury tolerance and mechanical response of the affected body region. When national automotive crash data from the United States were analysed by sex between 1998 and 2008, data revealed that the odds for a belt-restrained female driver to sustain severe injuries were 47% higher than those for a belt-restrained male driver involved in a comparable crash, after controlling for weight and body mass<sup>47</sup>. The subsequent introduction of a virtual female car crash dummy allowed mathematical simulations to account for the effect of acceleration on sex-specific biomechanics, highlighting the need to add a medium-sized female dummy model to regulatory safety testing<sup>48,49</sup>. Beyond automotive safety systems, the importance of anthropometric characteristics, such as the carrying angle of the elbow or the shape and size of the human knee, can be used to guide sex-specific design for artificial joints, limb prostheses and occupational protective gear<sup>50,51</sup>.

#### Reducing gender bias in AI

Alarming examples of algorithmic bias are well documented <sup>52</sup>. When translating gender-neutral language related to science, technology, engineering and mathematics (STEM) fields, Google Translate defaults to male pronouns <sup>53</sup>. When photographs depict a man in the kitchen, automated image captioning algorithms systematically misidentify the individual as a woman <sup>54</sup>. As Al becomes increasingly ubiquitous in everyday lives, such bias, if uncorrected, can amplify social inequities. Understanding how gender operates within the context of the algorithm helps researchers to make conscious decisions about how their work functions in society.

Since the Second World War, medical research has been submitted to stringent review processes aimed at protecting participants from harm. AI, which has the potential to influence human life at scale, has yet to be so carefully examined. Numerous groups have articulated 'principles' for human-centred AI. These include, most importantly, the UN Human Rights Framework that consists of internationally agreed upon human rights laws and standards, as well as the 'Asilomar AI Principles', 'AI at Google: Our Principles', 'Partnership on AI', and so on. What we lack are mechanisms for technologists to put these principles into practice. Here we delve into a few of such rapidly developing mechanisms for AI.

A first challenge in algorithmic bias is to identify when it is appropriate for an algorithm to use gender information. In some settings, such as the assignment of job ads, it might be desirable for the algorithm to explicitly ignore the gender of an individual as well as features such as weight, which may correlate with gender but are not directly related to job performance. In other applications, such as image/voice recognition, it might be desirable to leverage gender characteristics to achieve the best accuracy possible across all subpopulations. To date, there is no unified definition of algorithmic fairness <sup>55–57</sup>, and the best approach is to understand the nuances of each application domain, make transparent how algorithmic decision-making is deployed and appreciate how bias can arise <sup>58</sup>.

Training data are a source of potential bias in algorithms. Certain subpopulations, such as darker-skinned women, are often underrepresented in the data used to train machine-learning algorithms, and efforts are underway to collect more data from such groups<sup>2</sup>. To highlight the issue of underrepresented subpopulations in machine-learning data, researchers have designed 'nutrition labels' to capture metadata about how the dataset was collected and annotated 59-61. Useful metadata should summarize statistics on, for example, the sex, gender, ethnicity and geographical location of the participants in the dataset. In many machine-learning studies, the training labels are collected through crowdsourcing, and it is also useful to provide metadata about the demographics of crowd labellers.

Another approach to evaluate gender bias in algorithms is counterfactual analysis<sup>62</sup>. Consider Google Search, in which men are five times more likely than women to be offered ads for high-paying executive jobs<sup>63</sup>. The algorithm that decides which ad to show inputs features about the individual making the query and outputs a set of ads predicted to be relevant. The counterfactual would test the algorithm in silico by changing the gender of each individual in the data and then studying how predictions change. If simply changing an individual from 'woman' to 'man' systematically leads to higher paying job ads, then the predictor is-indeed-biased.

Work to debias word embeddings is another example of counterfactual analysis<sup>64</sup>. Word embeddings associate each English word with a vector of features so that the geometry between the feature vector captures semantic relations between the words. It is widely used in practice for applications such as sentiment analysis<sup>65</sup>, language translation<sup>66</sup> and analysis of electronic health records<sup>67</sup>. It has previously been shown that gender stereotypes-for example, men are more likely to be computer scientists—are manifested in the feature vectors of the corresponding words<sup>64</sup>. Whether this association between man and computer is problematic depends on the application of the features. To test for gender effects, gender-neutral word features were created. For each downstream application, counterfactual analysis can then be performed by running the application twice, once using the original word features, and once using the gender-neutral features. If the outcome changes, the algorithm is sensitive to gender. In some applications, such as job searches, it might be preferable to use gender-neutral features.

An alternative approach to quantify and reduce gender bias in algorithms is called multi-accuracy auditing<sup>68,69</sup>. In standard machine learning, the objective is to maximize the overall accuracy for the entire population, as represented by the training data. In multi-accuracy, the goal is to ensure that the algorithm achieves good performance not only in the aggregate but also for specific subpopulations—for example, 'elderly Asian man' or 'Native American woman'. The multi-accuracy auditor takes a complex machine-learning algorithm and systematically identifies whether the current algorithm makes more mistakes for any subpopulation. In a recent paper, the neural network used for facial recognition was audited and specific combinations of artificial neurons that responded to the images of darker-skinned women were identified that are responsible for the misclassifications<sup>70</sup>.

The auditor also suggests improvements when it identifies such biases<sup>71</sup>. Although achieving equal accuracy across all demographic groups may not always be feasible, these auditing techniques improve the transparency of the AI systems by quantifying how its performance varies across race, age, sex and intersections of these attributes.

These are only a few of the specific techniques computer scientists are developing to promote gender fairness in algorithms. Some, such as data checks, are relevant across all disciplines that amass and analyse big data. Others are specific to machine learning, which is now widely deployed across broad swathes of intellectual endeavours from the humanities to the social sciences, biomedicine and judicial systems. In all instances, it is important to be completely transparent where and for what purpose AI systems are used, and to characterize the behaviour of the system with respect to sex and gender<sup>72</sup>.

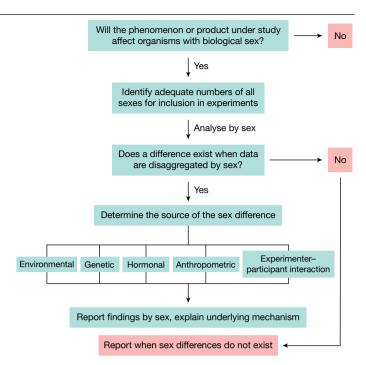


Fig. 2 | Sex analysis and reporting in science and engineering. This decision tree represents a cognitive process for analysing sex. A 'no' indicates no further analysis is necessary. A 'yes' suggests the next step that should be considered.

#### **Combatting stereotypes**

Analysing gender in software systems is one issue; configuring gender in hardware-such as social robots-is another, and the focus of this section. Until recently, robots were largely confined to factories. Most people never see or interact with these robots; they do not look, sound or behave like humans. But engineers are increasingly designing robots to assist humans as service robots in hospitals, elder care facilities, classrooms, homes, airports and hotels. The field of social human-robot interaction examines, among other things, when and how 'gendering' robots, virtual agents or chatbots might enhance usability while, at the same time, considering when and how to avoid oversimplifications that may reinforce potentially harmful gender stereotypes<sup>73</sup>.

Machines are, in principle, genderless. Gender, however, is a core social category in human impression formation that is readily applied to nonhuman entities<sup>74</sup>. Thus, users may consciously or unconsciously gender machines as a function of anthropomorphizing them, even when designers intend to create gender-neutral devices<sup>75–78</sup>.

Anthropomorphizing technologies may help users to engage more effectively with them, which poses the question as to whether there are benefits to tapping into the power of social stereotypes by building gender into virtual agents<sup>79-83</sup>, chatbots<sup>84</sup> or social robots<sup>11,85,86</sup>. For example, if roboticists deploy female carebots in female-typical roles, such as nursing, would users better comply with the robot's requests to take daily medication or to exercise? Does gendering robots or virtual agents facilitate interaction or boost objective outcomes such as performance<sup>11,80-91</sup>? Will personalizing robots or chatbots by gender increase consumer acceptance and, even, sales figures? Systematic empirical research is needed to address these open research issues.

What features lead humans to gender a robot? So far, experimental research designed to analyse robot gender has manipulated gender in a  $number of ways, including (1) \, by \, choosing \, a \, male \, or \, female \, name \, to \, label$ the robot<sup>87–92</sup>; (2) by colour-coding the robot<sup>93,94</sup>; (3) by manipulating visual indicators of gender (for example, face, hairstyle or lip colour 94,95); (4) by adding a male or female voice, or low or high pitch to simulate this, respectively 87-92,94,96,97; (5) by designing a gendered personality 87,98;

# **Perspective**

and (6) by deploying robots in gender-stereotypical domains, such as a male-voiced robot for security and a female-voiced robot in a healthcare role<sup>95</sup>. Other aspects, such as movements or gestures, that may potentially gender a robot still require empirical research<sup>85,86</sup>.

But there are dangers here. As soon as designers or users assign a gender to a machine, stereotypes follow. Designers of robots and AI do not simply create products that reflect our world, they also (perhaps unintentionally) reinforce and validate certain gender norms that are considered to be appropriate for men, women or gender nonconforming individuals  $^{11,73}$ .

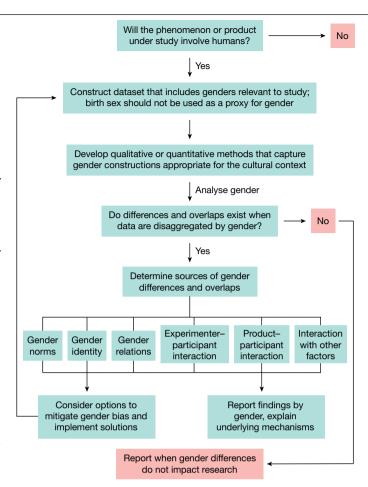
Eliciting gendered perceptions of technologies implies actively designing human gender biases, including binary constructions of gender as male or female, into machines. From a social psychological viewpoint, this can contribute to stereotypical gender norms in society95. Even though this might not seem relevant from an engineering point of view, social psychological research would suggest that a robot with a female appearance, for example, may perpetuate ideas of women as nurturing and communal, traits stereotypically associated with women<sup>95</sup>. Thus, a female robot may be deemed socially warm and particularly suitable for stereotypically female tasks, such as elderly care, or it might be openly sexualized and objectified as revealed in abusive commentary on video clips of female robots in recent qualitative research<sup>99</sup>. Similarly, virtual personal assistants with female names, voices and stereotypical, submissive behaviours, such as Siri or Alexa, represent heteronormative ideas about females and thereby indirectly contribute to the discrimination of women in society 100,101. An interesting development in this regard is the genderless voice, Q, which has recently been developed in Denmark to overcome such bias<sup>102</sup>.

There are many questions regarding these features. How, for example, do user attributes, such as age or gender, interact with different robot design features? How do robots enhance or harm real-world attitudes and behaviours related to social equality? How does robot gender elicit different responses across cultures? More experimental, laboratory and longitudinal field research is needed to test whether, and how, a machine's gendered, gender-diverse or gender-neutral appearance or behaviour influences human affect, cognition and behaviour. It is likely that even social robots designed to be genderless or gender neutral elicit gender attributions owing to the relatively automatic nature of anthropomorphizing humanoid robots. It is also likely that when potential end users are offered the option to select a digital assistant's gender, their choice will be driven by their own gender identity and gender-related attitudes and stereotypes. Addressing these research questions and issues remains important to shed light on the psychological, social and ethical implications of implicit or explicit design choices for novel technologies.

Developing technologies that enhance, or at least do not harm, social equality will require novel configurations of researchers. Much attention has been paid to the need for interdisciplinary research, consisting of humanists, legal experts, technologists and social scientists, especially in the fields of human-centred AI. The historical development of universities, however, has artificially separated human knowledge into disciplines over the course of the nineteenth and twentieth centuries that may not support current research needs. Research institutions now need to develop robust mechanisms to bring together social analysis and engineering in a way that rigorously addresses the emerging needs of society. 103.

#### Pathways to improving study design

To reach the full potential of sex and gender analysis for discovery and innovation, it is important to integrate sex and gender analysis, where relevant, into the design of research from the very beginning. Much of science and engineering research is path-dependent: once research has been designed, it becomes difficult to change. It is also important to understand that sex and gender are categories of analysis or variables (or



**Fig. 3** | **Gender analysis and reporting in science and engineering.** This decision tree represents a cognitive process for analysing gender. A 'no' indicates no further analysis is necessary. A 'yes' suggests the next step that should be considered.

controls) that need to be incorporated into the research process, but do not need be the main focus of the research. Nor will sex and gender analysis be relevant to all types of research. As the decision trees for analysing sex (Fig. 2) and gender (Fig. 3) indicate, in cases in which researchers have considered sex and/or gender but judge that this analysis is not relevant for a specific hypothesis, they may rule it out. Moreover, if researchers expect sex or gender to be important but find no significant differences, this may represent a result worthy of publication. Reporting cases in which sex or gender sameness, overlap or no difference is found may represent an important finding.

In this Perspective, we highlight the need and promise for designing sex and gender analysis into research through specific case studies and examples. From these, we extracted key considerations for analysing sex (Fig. 2) and gender (Fig. 3). These are generic recommendations that work across disciplines. However, more related studies are needed in the next five years. First, through interdisciplinary work, researchers need to sharpen and standardize generic approaches to sex and gender analysis that generalize across fields. Second, through discipline-specific work, researchers need to craft state-of-the-art analytics for study design and data analysis in their own subfields. The European Commission is currently funding an expert group that seeks to tailor sex and gender methods of analysis to field-specific protocols<sup>104</sup>.

#### **Future challenges**

We do not yet have results for sex and gender analysis in the physical sciences, such as basic chemistry, pure physics, geology or astronomy. Much work has analysed gender gaps in participation and gender bias

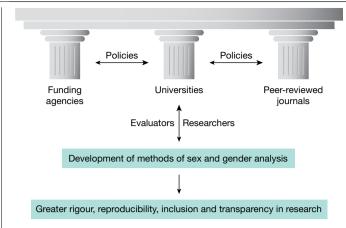


Fig. 4 | Three pillars of science and engineering infrastructure. To reap the benefits of sex and gender analysis, the pillars of science infrastructure must develop and implement coordinated policies.

in the culture of these fields, but attention has yet to turn to how the research itself may respond to gender analysis. As research in the physical sciences becomes more applied, sex and gender analysis become more relevant-for example, in the chemistry of aerosols, sex differences govern rates of inhalation and gender differences influence rates of exposure105.

Several methodological challenges remain for the field of sex and gender analysis itself. Although advances have been made in methods for analysing sex<sup>106</sup>, we lack non-invasive methods of sex determination in numerous non-model organisms, in which sexual morphological dimorphism is not easily detected. Technological advances through the development of genetic<sup>107</sup>, metabolomic<sup>108</sup> and endocrine<sup>3</sup> markers of organism sex are needed for non-model species at all stages of development, an endeavour that will be aided by the innovation and increased affordability of omics approaches. Attention will also need to be paid to the translation of evidence from animal species to humans as-in many cases-molecular sex differences observed in humans may not be mirrored in nonhuman mammals<sup>109</sup>.

Although sex as a biological variable in science and engineering is increasingly well understood<sup>110</sup>, the same cannot be said for gender as a cultural variable. Gender is complex and multidimensional (Facebook introduced 58 gender categories in 2014<sup>111</sup>) and applications in technical fields often require collaboration with social scientists to understand the relevant aspects of gender for specific projects. Even in health research, we lack systematic measures for assessing how gender relates to health because gender does not reduce easily to variables that can be manipulated statistically. Two recent studies have attempted to remedy this. The first used a binary gender index (masculinity versus femininity) constructed from seven variables and found that the incidence of recurrence and death 12 months after diagnosis of acute coronary syndrome in young adults was associated with gender and specifically not with biological sex<sup>112</sup>. A second study under development at Stanford University seeks to capture the multidimensionality of gender better by identifying theoretically robust gender-related variables relevant for health research. This study is based on US data, and new variables tailored to specific cultural settings need to be identified. Developing measures of gender is clearly an area for which more research is needed.

Other methodological challenges include going beyond the binaryfemale and male, women and men-in both sex and gender analysis. Take, for instance, the Gender API algorithm that allows social scientists to understand gender differences in research patterns. The algorithm identifies only binaries: female/male; woman/man. In the United States, 0.6% of the population-nearly 2 million people-identify as transgender 113, and more than 15 countries offer a third sex category on legal documents, birth certificates and passports. Research needs to keep pace with social change. Similarly, consider the lack of research that addresses how hermaphroditic animals respond to environmental change. In simultaneous hermaphrodites in which reproductively mature individuals have both male and female gametes, there is a need to consider the role of male or female tissues in determining the response of the whole organism. By contrast, in sequential hermaphrodites that change sex, there is a need to consider whether an organism responds as a female or a male to environmental stress during the sex change process, given that this process is dynamic, with behavioural, endocrine and genetic systems switching sex on markedly different timescales<sup>114</sup>.

Additional challenges include accounting for other social variables. such as age, race and geographical location, and how these intersect with sex and/or gender. Sex or gender cannot be isolated from other characteristics, and we need model systems and intersectional methods to understand these interrelationships 115. An intersectional approach in human research underscores the importance of unmasking and rectifying overlapping and interdependent systems of discrimination that are often built into knowledge, programs and policies. Benefits for global health, for example, will only be achieved when unbiased decision-making about resources takes into account the lived experiences of women and men with multiple identity characteristics who simultaneously suffer from race, class, education, economic and cultural power imbalance in accessing food and water, digital technology and healthcare services<sup>116</sup>.

#### **Science policy**

Policy is one driver of discovery and innovation that can enable sex and gender analysis in science and technology. To push forward rigorous sex and gender analysis, interlocking policies need to be implemented by three pillars of academic research: funding agencies, peer-reviewed journals and universities (Fig. 4).

Government-led funding agencies have taken the lead by asking applicants to explain how sex and gender analysis is relevant to their proposed research, or to explain that it is not (for a list of agencies and policies, see Supplementary Information section 1). The Canadian Institutes of Health Research showed robust uptake after mandating applicants to declare whether sex and/or gender were accounted for in proposals and to justify exclusion in 2010. Their evaluation revealed that from 2010–2011 the proportion of funded proposals incorporating sex and/ or gender analysis nearly doubled 117,118.

The second pillar, peer-reviewed journals, have developed editorial policies advocating for sex or gender analysis to ensure excellence in papers selected for publication (for a list of journals and policies, see Supplementary Information section 2). Uptake has been swift in health and medicine. The Lancet, for example, adopted such guidelines in 2016, followed quickly by the International Committee of Medical Journal Editors<sup>119</sup>. The Structured, Transparent, Accessible Reporting (STAR) methods of Cell Press have required transparent reporting of the sex distribution of donor cells, also since 2016. Importantly, the widely adopted Sex and Gender Equity in Reporting (SAGER) guidelines recommend that data be disaggregated by both sex and gender<sup>120</sup>. Although biomedical journals have moved rapidly, we are not aware of any engineering or computer science conferences or journals with such guidelines.

Pillars one and two need the support of a third pillar: universities. Both funding agencies and journals may have policies in place, but researchers and evaluators by and large lack expertise in sex and gender analysis. The European Commission, which has had policies in place since 2014, found that fewer than expected funded research proposals incorporated sex and gender analysis and has correlated this low proportion to an 'absence of training on gender issues'121. Similarly, an analysis of animal research in the neurosciences showed that in 2014 only about 14% of peer-reviewed articles considered sex as a biological variable 122.

Universities need to step up and incorporate sex and gender analysis as a conceptual tool into science and engineering curricula. Numerous universities offer gender analysis in the humanities and social sciences,

### **Perspective**

but not in core natural science and engineering courses. Efforts have been made in medicine—the Charité in Berlin, Germany, for instance, has successfully integrated sex and gender analysis throughout all six years of medical training from early basic science to later clinical modules<sup>123</sup>. However, this is a rare example, and universities must do more to prepare the scientific workforce for the future.

Several initiatives have endeavoured to fill this gap. Gendered Innovations—a global, collaborative project initiated from Stanford University in 2009 and supported by the European Commission and the US National Science Foundation—has developed practical methods of sex and gender analysis for natural scientists and engineers, and provides case studies as concrete illustrations of how sex and gender analysis lead to discovery and innovation (https://genderedinnovations.stanford.edu/). The WHO (World Health Organization) has developed a gender-responsive assessment tool 124. The Organization for the Study of Sex Differences (https://www.ossdweb.org/) has advanced sex and gender analysis methods for the life and health sciences. The Canadian Institutes of Health Research have developed online training modules for integrating sex and gender analysis into biomedical research 125. These initiatives should now be mainstreamed into university education.

Much work remains to be done to systematically integrate sex and gender analysis into relevant domains of science and technology—from strategic considerations for establishing research priorities to guidelines for establishing best practices in formulating research questions, designing methodologies and interpreting data. To make real progress in the next decade, researchers, funding agencies, peer-reviewed journals and universities need to coordinate efforts to develop and standardize methods of sex and gender analysis.

But eyes have been opened, and by integrating sex and gender analysis into their work, researchers can enhance excellence and social responsibility in science and engineering.

- Conforti, F. et al. Cancer immunotherapy efficacy and patients' sex: a systematic review and meta-analysis. Lancet Oncol. 19, 737–746 (2018).
- Buolamwini, J. & Gebru, T. Gender shades: intersectional accuracy disparities in commercial gender classification. Proc. Mach. Learn. Res. 81, 77–91 (2018).
   This paper demonstrates that commercial gender-identification algorithms misclassify darker-skinned females at a higher rate than the rest of the population, which is an example of how algorithmic bias intersects with gender and race.
- Jensen, M. P. et al. Environmental warming and feminization of one of the largest sea turtle populations in the world. Curr. Biol. 28, 154–159 (2018).
- Parker, L. M. et al. Ocean acidification but not warming alters sex determination in the Sydney rock oyster, Saccostrea glomerata. Proc. R. Soc. Lond. B 285, 20172869 (2018).
- 5. Baker, M. 1,500 scientists lift the lid on reproducibility. Nature 533, 452-454 (2016).
- American Society of Cell Biology. Member survey on reproducibility. http://www.ascb. org/wp-content/uploads/2015/11/final-survey-results-without-Q11.pdf (2014).
- Shah, K., McCormack, C. E. & Bradbury, N. A. Do you know the sex of your cells? Am. J. Physiol. Cell Physiol. 306, C3–C18 (2014). This review demonstrates that the sex of cells used in experiments can influence the biology of the cell and provides a table outlining the sex of cell lines that have appeared in Am. J. Physiol. Cell Physiol. in recent years.
- Potluri, T., Engle, K., Fink, A. L., Vom Steeg, L. G. & Klein, S. L. Sex reporting in preclinical microbiological and immunological research. mBio 8, e01868-17 (2017).
- Ellis, R. P. et al. Does sex really matter? Explaining intraspecies variation in ocean acidification responses. Biol. Lett. 13, 20160761 (2017).
- Bahamonde, P. A., Feswick, A., Isaacs, M. A., Munkittrick, K. R. & Martyniuk, C. J. Defining the role of omics in assessing ecosystem health: perspectives from the Canadian environmental monitoring program. *Environ. Toxicol. Chem.* 35, 20–35 (2016).
- Nomura, T. Robots and gender. Gend. Genome 1, 18–26 (2017).
   This paper presents an overview of research on the role of sex and gender in the context of human-robot interaction research and summarizes the cutting-edge research in this area from an engineering perspective.
- 12. Cripps, G., Flynn, K. J. & Lindeque, P. K. Ocean acidification affects the phyto-zoo plankton trophic transfer efficiency. PLoS ONE 11, e0151739 (2016). This paper discusses how sex analysis reveals significant effects of ocean acidification on the respiration rate in marine copepods, with males and females showing differential baseline respiration rates and responding to increased levels of CO<sub>2</sub> in opposing directions.
- Li, J., Ju, W. & Reeves, B. Touching a mechanical body: tactile contact with intimate parts of a humanoid robot is physiologically arousing. J. Hum. Robot Interact. 6, 118–130 (2017).
- Suvilehto, J. T., Glerean, E., Dunbar, R. I. M., Hari, R. & Nummenmaa, L. Topography of social touching depends on emotional bonds between humans. *Proc. Natl Acad. Sci. USA* 112, 13811–13816 (2015).
- Becker, J. B., Prendergast, B. J. & Liang, J. W. Female rats are not more variable than male rats: a meta-analysis of neuroscience studies. *Biol. Sex Differ.* 7, 34 (2016).

- Prendergast, B. J., Onishi, K. G. & Zucker, I. Female mice liberated for inclusion in neuroscience and biomedical research. Neurosci. Biobehav. Rev. 40, 1–5 (2014).
- Itoh, Y. & Arnold, A. P. Are females more variable than males in gene expression? Metaanalysis of microarray datasets. Biol. Sex Differ. 6, 18 (2015).
- Romanov, N. et al. Disentangling genetic and environmental effects on the proteotypes of individuals. Cell 177, 1308–1318 (2019).
- Beery, A. K. Inclusion of females does not increase variability in rodent research studies. Curr. Opin. Behav. Sci. 23, 143–149 (2018).
- Buch, T. et al. Benefits of a factorial design focusing on inclusion of female and male animals in one experiment. J. Mol. Med. 97, 871–877 (2019).
- Miller, L. R. et al. Considering sex as a biological variable in preclinical research. FASEB J. 31, 29–34 (2017).
- Maures, T. J. et al. Males shorten the life span of C. elegans hermaphrodites via secreted compounds. Science 343, 541–544 (2014).
- Chapman, C. D., Benedict, C. & Schi
  öth, H. B. Experimenter gender and replicability in science. Sci. Adv. 4, e1701427 (2018).
- Sorge, R. E. et al. Olfactory exposure to males, including men, causes stress and related analgesia in rodents. Nat. Methods 11, 629–632 (2014).
- Nielsen, M. W., Andersen, J. P., Schiebinger, L. & Schneider, J. W. One and a half million medical papers reveal a link between author gender and attention to gender and sex analysis. Nat. Hum. Behav. 1, 791-796 (2017).
- Sugimoto, C. R., Ahn, Y. Y., Smith, E., Macaluso, B. & Larivière, V. Factors affecting sexrelated reporting in medical research: a cross-disciplinary bibliometric analysis. *Lancet* 393, 550–559 (2019).
- 27. Ainsworth, C. Sex redefined, Nature 518, 288-291 (2015).
- Bachtrog, D. et al. Sex determination: why so many ways of doing it? PLoS Biol. 12, e1001899 (2014).
- Ospina-Álvarez, N. & Piferrer, F. Temperature-dependent sex determination in fish revisited: prevalence, a single sex ratio response pattern, and possible effects of climate change. PLoS ONE 3, e2837 (2008).
- Munday, P. L., Buston, P. M. & Warner, R. R. Diversity and flexibility of sex-change strategies in animals. *Trends Ecol. Evol.* 21, 89–95 (2006).
- Parker, K. A direct method for estimating northern anchovy, Engraulis mordax, spawning biomass. Fish Bull. 78, 541–544 (1980).
- 32. Barneche, D. R., Robertson, D. R., White, C. R. & Marshall, D. J. Fish reproductive-energy output increases disproportionately with body size. *Science* **360**, 642–645 (2018).
- Shang, E. H. H., Yu, R. M. K. & Wu, R. S. S. Hypoxia affects sex differentiation and development, leading to a male-dominated population in zebrafish (*Danio rerio*). Environ. Sci. Technol. 40, 3118–3122 (2006).
- Oldfield, R. G. Genetic, abiotic and social influences on sex differentiation and the evolution of sequential hermaphroditism. Fish Fish. 6, 93–110 (2005).
- Kindsvater, H. K., Reynolds, J. D., Sadovy de Mitcheson, Y. & Mangel, M. Selectivity matters: rules of thumb for management of plate-sized, sex-changing fish in the live reef food fish trade. Fish Fish. 18. 821–836 (2017).
- Benvenuto, C., Coscia, I., Chopelet, J., Sala-Bozano, M. & Mariani, S. Ecological and evolutionary consequences of alternative sex-change pathways in fish. Sci. Rep. 7, 9084 (2012)

This paper demonstrates that the direction of sex change in sequentially hermaphroditic fish is a critical variable determining demographic stability and resilience, with considerable implications concerning the resilience of these populations to anthropogenic disturbances, such as overfishing.

- Casas, L. et al. Sex change in clownfish: molecular insights from transcriptome analysis. Sci. Rep. 6, 35461 (2016).
- Labonté, B. et al. Sex-specific transcriptional signatures in human depression. Nat. Med. 23, 1102–1111 (2017).
- Sorge, R. E. et al. Different immune cells mediate mechanical pain hypersensitivity in male and female mice. Nat. Neurosci. 18, 1081–1083 (2015).
- Tannenbaum, C. & Day, D. Age and sex in drug development and testing for adults. Pharmacol. Res. 121, 83-93 (2017).
- Juul, K. V., Klein, B. M., Sandström, R., Erichsen, L. & Nørgaard, J. P. Gender difference in antidiuretic response to desmopressin. Am. J. Physiol. Renal Physiol. 300, F1116–F1122 (2011)
- Giefing-Kröll, C., Berger, P., Lepperdinger, G. & Grubeck-Loebenstein, B. How sex and age
  affect immune responses, susceptibility to infections, and response to vaccination. Aging
  Cell 14, 309–321 (2015).
- Gao, F. et al. XWAS: a software toolset for genetic data analysis and association studies of the X chromosome. J. Hered. 106, 666–671 (2015).
- 44. Wise, A. L., Gyi, L. & Manolio, T. A. Exclusion: toward integrating the X chromosome in genome-wide association analyses. *Am. J. Hum. Genet.* **92**, 643–647 (2013).
- Khramtsova, E. A., Davis, L. K. & Stranger, B. E. The role of sex in the genomics of human complex traits. *Nat. Rev. Genet.* 20, 173-190 (2019).
- This review discusses new techniques for sex analysis in genome-wide association studies, epigenetic studies and X-chromosome inactivation.
- 46. Schiebinger, L. Women and Gender in Science and Technology (Routledge, 2014).
- Bose, D., Segui-Gomez, M. & Crandall, J. R. Vulnerability of female drivers involved in motor vehicle crashes: an analysis of US population at risk. Am. J. Public Health 101, 2368–2373 (2011).
- Linder, A., Holmqvist, K. & Svensson, M. Y. Average male and female virtual dummy model (BioRID and EvaRID) simulations with two seat concepts in the Euro NCAP low severity rear impact test configuration. Accid. Anal. Prev. 114, 62–70 (2018)
- Linder, A. & Svedberg, W. Review of average sized male and female occupant models in European regulatory safety assessment tests and European laws: gaps and bridging suggestions. Accid. Anal. Prev. 127, 156–162 (2019).
- Falys, C. G., Schutkowski, H. & Weston, D. A. The distal humerus—a blind test of Rogers' sexing technique using a documented skeletal collection. J. Forensic Sci. 50, JFS2005171 (2005).

- Conley, S., Rosenberg, A. & Crowninshield, R. The female knee: anatomic variations. J. Am. Acad. Orthop. Surg. 15, S31-S36 (2007)
- 52. Zou, J. & Schiebinger, L. Al can be sexist and racist — it's time to make it fair. Nature 559, 324-326 (2018).
- Prates, M., Avelar, P. & Lamb, L. Assessing gender bias in machine translation—a case study with Google translate. Neural Comput. Appl. https://doi.org/10.1007/s00521-019-
- 54. Zhao, J., Wang, T., Yatskar, M., Ordonez, V. & Chang, K. W. Men also like shopping: reducing gender bias amplification using corpus-level constraints. In Proc. 2017 Conference on Empirical Methods in Natural Language Processing (eds Palmer, M. et al.) 2979-2989 (ACL, 2017).
- Corbett-Davies, S. & Goel, S. The measure and mismeasure of fairness: a critical review of fair machine learning. Preprint at https://arxiv.org/abs/1808.00023 (2018).
- 56. Dwork, C., Hardt, M., Pitassi, T., Reingold, O. & Zemel, R. Fairness through awareness. In Proc. 3rd Innovations in Theoretical Computer Science Conference 214-226 (ACM, 2012). This paper develops a mathematical framework to study the fairness notion that similar individuals should be treated similarly by algorithms.
- 57. Zemel, R., Wu, Y., Swersky, K., Pitassi, T. & Dwork, C. Learning fair representations, Proc. Mach. Learn. Res. 28, 325-333 (2013).
- Barocas, S. & Selbst, A. D. Big data's disparate impact. Calif. Law Rev. 104. 671-732 (2016). 58. This paper studies what it means for algorithms to create disparate impact and to discriminate through the lens of American antidiscrimination law, and discusses the legal challenges involved
- 59. Gebru, T. et al. Datasheets for datasets. Preprint at https://arxiv.org/abs/1803.09010 (2018).
- 60. Holland, S., Hosny, A., Newman, S., Joseph, J. & Chmielinski, K. The dataset nutrition label: a framework to drive higher data quality standards. Preprint at https://arxiv.org/ abs/1805.03677 (2018)
- 61. Bender, E. M. & Friedman, B. Data statements for NLP: toward mitigating system bias and enabling better science. Preprint at https://openreview.net/forum?id=By4oPeX9f (2018).
- Kusner, M., Loftus, J., Russell, C. & Silva, R. Counterfactual fairness. Adv. Neural Inf. Process. Syst. 30, 4066-4076 (2017).
- 63. Datta, A., Tschantz, M. C. & Datta, A. Automated experiments on ad privacy settings. Proc. Privacy Enhancing Technol. 2015, 92-112 (2015).
- 64. Bolukbasi, T., Chang, K. W., Zou, J., Saligrama, V. & Kalai, A. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Adv. Neural Inf. Process. Syst. 29, 4349-4357 (2016).
- Maas, A. L. et al. Learning word vectors for sentiment analysis. In Proc. 49th Annual 65. Meeting of the Association for Computational Linguistics: Human Language Technologies (eds Matsumoto, Y. & Mihalcea, R.) 142-150 (ACL, 2011).
- Zou, W. Y., Socher, R., Cer, D. & Manning, C. D. Bilingual word embeddings for phrase-66. based machine translation. In Proc. 2013 Conference on Empirical Methods in Natural Language Processing (eds Yarowsky, D. et al.) 1393-1398 (ACL, 2013).
- Nie, A. et al. DeepTag: inferring diagnoses from veterinary clinical notes. NPJ Digit. Med. 1, 60 (2018).
- 68 Hébert-Johnson, U., Kim, M. P., Reingold, O. & Rothblum, G. N. Multicalibration: calibration for the (computationally-identifiable) masses. Proc. Mach. Learn. Res. 80, 1939-1948 (2018)
- Kearns, M., Neel, S., Roth, A. & Wu, Z. S. Preventing fairness gerrymandering: auditing and learning for subgroup fairness. Proc. Mach. Learn. Res. 80, 2564-2572 (2018).
- 70. Kim, M. P., Ghorbani, A. & Zou, J. Multiaccuracy: black-box post-processing for fairness in classification. Preprint at https://arxiv.org/abs/1805.12317 (2018).
- Doshi-Velez, F. & Kim, B. Towards a rigorous science of interpretable machine learning. Preprint at https://arxiv.org/abs/1702.08608 (2017).
- West, S. M., Whittaker, M. & Crawford, K. Discriminating systems: gender, race and power in Al. Al Now Institute https://ainowinstitute.org/discriminatingsystems.html
- Wang, Y. & Young, J. E. Beyond pink and blue: gendered attitudes towards robots in society. In Proc. 2nd Conference on Gender and IT Appropriation 49-54 (European Society for Socially Embedded Technologies, 2014).
- Fiske, S. T. in The Handbook of Social Psychology (eds Gilbert, D. T. et al.) 357-411 (McGraw-Hill, 1998).
- 75 Epley, N., Waytz, A. & Cacioppo, J. T. On seeing human: a three-factor theory of anthropomorphism. Psychol. Rev. 114, 864-886 (2007).

#### This paper presents a psychological model of anthropomorphism, specifying three factors that elicit anthropomorphic inferences about nonhuman entities

- von Zitzewitz, J., Boesch, P. M., Wolf, P. & Riener, R. Quantifying the human likeness of a 76. humanoid robot, Int. J. Soc. Robot, 5, 263-276 (2013).
- Lemaignan, S., Fink, J., Dillenbourg, P. & Braboszcz, C. The cognitive correlates of anthropomorphism. In 2014 Human-Robot Interaction Conference, Workshop on "HRI: A Bridge between Robotics and Neuroscience" (2014).
- 78 Nass, C. & Moon, Y. Machines and mindlessness: social responses to computers, J. Soc. Issues 56, 81-103 (2000).
- Gulz, A. & Haake, M. in Gender Issues in Learning and Working with Information Technology: Social Constructs and Cultural Contexts (eds Goodman, S. et al.) 113-132 (IGI Global, 2010)
- Krämer, N. C. et al. Closing the gender gap in STEM with friendly male instructors? On the effects of rapport behavior and gender of a virtual agent in an instructional interaction. Comput. Educ. 99, 1-13 (2016).
- Baylor, A. L. The design of motivational agents and avatars. Educ. Technol. Res. Dev. 59,
- Arroyo, I., Woolf, B. P., Royer, J. M. & Tai, M. Affective gendered learning companions. In Proc. 2009 Conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling (eds Dimitrova, V. et al.) 41-48 (IOS Press, 2009).
- Baylor, A. L. Promoting motivation with virtual agents and avatars: role of visual presence and appearance. Phil. Trans. R. Soc. B 364, 3559-3565 (2009).

- McDonnell, M. & Baxter, D. Chatbots and gender stereotyping. Interact, Comput. 31, 116-121 (2019)
- 85. Alesich, S. & Rigby, M. Gendered robots: implications for our humanoid future. IEEE Technol. Soc. Mag. 36, 50-59 (2017).
- Søraa, R. A. Mechanical genders: how do humans gender robots? Gend. Technol. Dev. 21, 99-115 (2017).
- Kraus, M., Kraus, J., Baumann, M. & Minker, W. Effects of gender stereotypes on trust and likability in spoken human-robot interaction. In Proc. 11th International Conference on Language Resources and Evaluation (eds Calzolari, N. et al.) 112-118 (European Language Resources Association, 2018).
- Crowell, C. R., Scheutz, M., Schermerhorn, P. & Villano, M. Gendered voice and robot entities: perceptions and reactions of male and female subjects. In Proc. 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems 3735-3741 (IEEE, 2009).
- Alexander, E., Bank, C., Yang, J. J., Hayes, B. & Scassellati, B. Asking for help from a gendered robot. In Proc. 36th Annual Meeting of the Cognitive Science Society 2333-2338 (Curran Associates, 2014).
- Kuchenbrandt, D., Häring, M., Eichberg, J., Eyssel, F. & André, E. Keep an eye on the task! How gender typicality of tasks influence human-robot interactions, Int. J. Soc. Robot. 6. 417-427 (2014).
- Reich-Stiebert, N. & Eyssel, F. (Ir)relevance of gender? On the influence of gender stereotypes on learning with a robot. In Proc. 2017 ACM/IEEE International Conference on Human-Robot Interaction (eds Mutlu, B. & Tscheligi, M.) 166-176 (ACM 2017)
- Tay, B., Jung, Y. & Park, T. When stereotypes meet robots: the double-edge sword of robot gender and personality in human-robot interaction. Comput. Human Behav. 38, 75-84 (2014).
- Jung, E. H., Waddell, T. F. & Sundar, S. S. Feminizing robots: user responses to gender cues on robot body and screen. In Proc. 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems 3107-3113 (ACM, 2016).
- Powers, A. et al. Eliciting information from people with a gendered humanoid robot. In Proc. 2005 IEEE International Workshop on Robots and Human Interactive Communication 158-163 (IEEE, 2005)
- Eyssel, F. & Hegel, F. (S)he's got the look: gender stereotyping of robots. J. Appl. Soc. Psychol. 42, 2213-2230 (2012).
- Siegel, M., Breazeal, C. & Norton, M. I. Persuasive robotics: the influence of robot gender on human behavior. In Proc. 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems 2563-2568 (IEEE, 2009).
- Eyssel, F., Kuchenbrandt, D., Hegel, F. & de Ruiter, L. Activating elicited agent knowledge: how robot and user features shape the perception of social robots. In Proc. 21st IEEE international Symposium on Robot and Human Interactive Communication 851-857 (IEEE,
- Kittmann, R. et al. Let me introduce myself: I am Care-O-bot 4, a gentleman robot. In Proc. 98 Mensch und Computer 2015 223-232 (De Gruyter Oldenbourg, 2015).
- Strait M.K. Aquillon C. Contreras V. & Garcia N. The public's perception of humanlike robots: online social commentary reflects an appearance-based uncanny valley, a general fear of a "Technology Takeover", and the unabashed sexualization of femalegendered robots. In 26th IEEE International Symposium on Robot and Human Interactive Communication 1418-1423 (IEEE, 2017).
- 100. Loideain, N. N. & Adams, R. From Alexa to Siri and the GDPR; the gendering of virtual personal assistants and the role of EU data protection law. King's College London Dickson Poon School of Law Legal Studies Research Paper Series https://doi.org/10.2139/ ssrn.3281807 (King's College London, 2018).
- West, M., Kraut, R. & Chew, H. E. I'd Blush If I Could: Closing Gender Divides in Digital Skills 101. through Education https://unesdoc.unesco.org/ark:/48223/pf0000367416 (EQUALS/ UNESCO, 2019).
- 102. Meet, Q. The First Genderless Voice https://www.genderlessvoice.com/ (2019).
- 103. Nielsen, M. W., Bloch, C. W. & Schiebinger, L. Making gender diversity work for scientific discovery and innovation. Nat. Hum. Behav. 2, 726-734 (2018).
- 104. European Commission. H2020 Expert Group to Update and Expand "Gendered Innovations/Innovation through Gender" http://ec.europa.eu/transparency/regexpert/ index.cfm?do=groupDetail.groupDetail&groupID=3601&NewSearch=1&NewSearch=1
- 105. Lorenz, C. et al. Nanosized aerosols from consumer sprays; experimental analysis and exposure modeling for four commercial products. J. Nanopart. Res. 13, 3377-3391
- 106. Mauvais-Jarvis, F., Arnold, A. P. & Reue, K. A guide for the design of pre-clinical studies on sex differences in metabolism. Cell Metab. 25, 1216-1230 (2017).
- 107. Hines, A. et al. Comparison of histological, genetic, metabolomics, and lipid-based methods for sex determination in marine mussels, Anal. Biochem. 369, 175-186 (2007).
- 108. Ellis, R. P. et al. <sup>1</sup>H NMR metabolomics reveals contrasting response by male and female mussels exposed to reduced seawater pH, increased temperature, and a pathogen. Environ. Sci. Technol. 48, 7044-7052 (2014).
- 109. Naqvi, S. et al. Conservation, acquisition, and functional impact of sex-biased gene expression in mammals. Science 365, eaaw7317 (2019).
- 110. Clayton, J. A. Studying both sexes: a guiding principle for biomedicine. FASEB J. 30, 519-524 (2016)
- Bivens, R. The gender binary will not be deprogrammed: ten years of coding gender on Facebook. New Media Soc. 19, 880-898 (2017).
- Pelletier, R., Ditto, B. & Pilote, L. A composite measure of gender and its association with risk factors in patients with premature acute coronary syndrome, Psychosom, Med. 77.
- Flores, A. R., Herman, J. L., Gates, G. G. & Brown, T. N. T. How Many Adults Identify as Transgender in the United States.https://williamsinstitute.law.ucla.edu/research/howmany-adults-identify-as-transgender-in-the-united-states/ (The Williams Institute, 2016).
- 114. Lamm, M. S., Liu, H., Gemmell, N. J. & Godwin, J. R. The need for speed: neuroendocrine regulation of socially-controlled sex change. Integr. Comp. Biol. 55, 307-322 (2015).

### **Perspective**

- Rice, C., Harrison, E. & Friedman, M. Doing justice to intersectionality in research. Cult. Stud. Crit. Methodol. 19, 409-420 (2019)
- Heise, L. et al. Gender inequality and restrictive gender norms: framing the challenges to health. Lancet 393, 2440-2454 (2019).
- Johnson, J., Sharman, Z., Vissandjée, B. & Stewart, D. E. Does a change in health research funding policy related to the integration of sex and gender have an impact? PLoS ONE 9,
- 118. Duchesne, A., Tannenbaum, C. & Einstein, G. Funding agency mechanisms to increase sex and gender analysis. Lancet 389, 699 (2017).
- Schiebinger, L., Leopold, S. S. & Miller, V. M. Editorial policies for sex and gender analysis. Lancet 388, 2841-2842 (2016).
- 120. Heidari, S., Babor, T. F., De Castro, P., Tort, S. & Curno, M. Sex and gender equity in research: rationale for the SAGER guidelines and recommended use. Res. Integr. Peer Rev. 1, 2 (2016).
- 121. Directorate-General for Research and Innovation. Interim Evaluation: Gender Equality as a Crosscutting Issue in Horizon 2020 (European Commission, 2017.)
- 122. Will, T. R. et al. Problems and progress regarding sex bias and omission in neuroscience research, eNeuro 4, ENEURO.0278-17.2017 (2017).
- 123. Ludwig, S. et al., A successful strategy to integrate sex and gender medicine into a newly developed medical curriculum, J. Womens Health 24, 996-1005 (2015).
- 124. World Health Organization. Gender Mainstreaming for Health Managers: A Practical Approach https://www.who.int/gender-equity-rights/knowledge/health\_managers\_ auide/en (2011).
- Canadian Institutes of Health Research. Online Training Modules: Integrating Sex & Gender in Health Research http://www.cihr-irsc.gc.ca/e/49347.html (2017).
- 126. Arboleda, V. A., Sandberg, D. E. & Vilain, E. DSDs: genetics, underlying pathologies and psychosexual differentiation. Nat. Rev. Endocrinol. 10, 603-615 (2014).
- 127. Hughes, I. A., Houk, C., Ahmed, S. F. & Lee, P. A. Consensus statement on management of intersex disorders. J. Pediatr. Urol. 2, 148-162 (2006).
- 128. Jarne, P. & Auld, J. R. Animals mix it up too: the distribution of self-fertilization among hermaphroditic animals. Evolution 60, 1816-1824 (2006).
- 129. Schiebinger, L. et al. Gender http://genderedinnovations.stanford.edu/terms/gender. html (2011-2019).

130. Boerner, K. E. et al. Conceptual complexity of gender and its relevance to pain. Pain 159, 2137-2141 (2018).

Acknowledgements We thank H. F. LeBlanc for her assistance. R.E. acknowledges financial support from a NERC Industrial Innovation Fellowship (NE/R013241/1). J.Z. is supported by a Chan-Zuckerberg Investigator Award and NIH P30AG059307. The views expressed do not necessarily reflect those of the Canadian Institutes of Health Research or the Canadian

Author contributions L.S. conceptualized the paper and invited R.E., F.E., C.T. and J.Z. to collaborate, L.S. and C.T. structured and drafted the article, R.E. wrote the marine science section, F.F. wrote the social robots section, L.S. wrote the introductory and policy sections C.T. wrote the health and medicine sections and J.Z. wrote the machine-learning section, All authors commented on and revised the paper, R.E. conceived and developed Fig. 1, C.T. conceived and developed Figs. 2 and 3, and contributed to Fig. 1, L.S. contributed to Fig. 3 and developed Fig. 4.

Competing interests The authors declare no competing interests.

#### Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-

Correspondence and requests for materials should be addressed to L.S.

Peer review information Nature thanks Simon Beggs, Cynthia Breazeal, Jayne Danska, Reshma Jagsi and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Reprints and permissions information is available at http://www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© Springer Nature Limited 2019

# A small proton charge radius from an electron-proton scattering experiment

https://doi.org/10.1038/s41586-019-1721-2

Received: 17 June 2019

Accepted: 19 September 2019

Published online: 6 November 2019

W. Xiong<sup>1</sup>, A. Gasparian<sup>2\*</sup>, H. Gao<sup>1</sup>, D. Dutta<sup>3\*</sup>, M. Khandaker<sup>4</sup>, N. Liyanage<sup>5</sup>, E. Pasyuk<sup>6</sup>, C. Peng<sup>1</sup>, X. Bai<sup>5</sup>, L. Ye<sup>3</sup>, K. Gnanvo<sup>5</sup>, C. Gu<sup>1</sup>, M. Levillain<sup>2</sup>, X. Yan<sup>1</sup>, D. W. Higinbotham<sup>6</sup>, M. Meziane<sup>1</sup>, Z. Ye<sup>1,7</sup>, K. Adhikari<sup>3</sup>, B. Aljawrneh<sup>2</sup>, H. Bhatt<sup>3</sup>, D. Bhetuwal<sup>3</sup>, J. Brock<sup>6</sup>, V. Burkert<sup>6</sup>, C. Carlin<sup>6</sup>, A. Deur<sup>6</sup>, D. Di<sup>5</sup>, J. Dunne<sup>3</sup>, P. Ekanayaka<sup>3</sup>, L. El-Fassi<sup>3</sup>, B. Emmich<sup>3</sup>, L. Gan<sup>8</sup>, O. Glamazdin<sup>9</sup>, M. L. Kabir<sup>3</sup>, A. Karki<sup>3</sup>, C. Keith<sup>6</sup>, S. Kowalski<sup>10</sup>, V. Lagerquist<sup>11</sup>, I. Larin<sup>12,13</sup>, T. Liu<sup>1</sup>, A. Liyanage<sup>14</sup>, J. Maxwell<sup>6</sup>, D. Meekins<sup>6</sup>, S. J. Nazeer<sup>14</sup>, V. Nelyubin<sup>5</sup>, H. Nguyen<sup>5</sup>, R. Pedroni<sup>2</sup>, C. Perdrisat<sup>15</sup>, J. Pierce<sup>6</sup>, V. Punjabi<sup>16</sup>, M. Shabestari<sup>3</sup>, A. Shahinyan<sup>17</sup>, R. Silwal<sup>10</sup>, S. Stepanyan<sup>6</sup>, A. Subedi<sup>3</sup>, V. V. Tarasov<sup>12</sup>, N. Ton<sup>5</sup>, Y. Zhang<sup>1</sup> & Z. W. Zhao<sup>1</sup>

Elastic electron–proton scattering (e–p) and the spectroscopy of hydrogen atoms are the two methods traditionally used to determine the proton charge radius,  $r_{\rm p}$ . In 2010, a new method using muonic hydrogen atoms¹ found a substantial discrepancy compared with previous results<sup>2</sup>, which became known as the 'proton radius puzzle'. Despite experimental and theoretical efforts, the puzzle remains unresolved. In fact, there is a discrepancy between the two most recent spectroscopic measurements conducted on ordinary hydrogen<sup>3,4</sup>. Here we report on the proton charge radius experiment at Jefferson Laboratory (PRad), a high-precision e-p experiment that was established after the discrepancy was identified. We used a magnetic-spectrometerfree method along with a windowless hydrogen gas target, which overcame several limitations of previous e-p experiments and enabled measurements at very small forward-scattering angles. Our result,  $r_D = 0.831 \pm 0.007_{\text{stat}} \pm 0.012_{\text{syst}}$  femtometres, is smaller than the most recent high-precision e-p measurement<sup>5</sup> and 2.7 standard deviations smaller than the average of all e-p experimental results<sup>6</sup>. The smaller  $r_p$  we have now measured supports the value found by two previous muonic hydrogen experiments<sup>1,7</sup>. In addition, our finding agrees with the revised value (announced in 2019) for the Rydberg constant<sup>8</sup>—one of the most accurately evaluated fundamental constants in physics.

The proton is the dominant component of visible matter in the Universe. Consequently, determining the proton's basic properties—such as its root-mean-square charge radius,  $r_p$  — is of interest in its own right. Accurate knowledge of  $r_n$  is also important for the precise determination of other fundamental constants, such as the Rydberg constant  $(R_{\infty})^2$ . The value of  $r_p$  is also required for precise calculations of the energy levels and transition energies of the hydrogen atom—for example, the Lamb shift. In muonic hydrogen (µH atoms), in which the electron in the H atom is replaced by a 'heavier electron' (a muon), the extended proton charge distribution changes the Lamb shift by as much as  $2\%^1$ . The first-principles calculation of  $r_p$  from the accepted theory of the strong interaction (quantum chromodynamics, QCD), is notoriously challenging and currently cannot reach the accuracy demanded by experiments, but lattice QCD calculations are on the cusp of becoming precise enough to be tested experimentally<sup>9</sup>. Therefore, the precise measurement of  $r_p$  is not only critical for addressing the proton radius puzzle but also important for determining certain fundamental constants of physics and testing lattice QCD.

Prior to 2010 the two methods used to measure  $r_p$  were ep  $\rightarrow$  ep elastic scattering measurements, in which the slope of the extracted proton (p) electric (E) form factor,  $G_{\rm F}^{\rm p}$ , as the four-momentum transfer squared  $(Q^2)$  approaches zero, is proportional to  $r_{p'}^2$  and Lamb shift (spectroscopy) measurements of ordinary H atoms, which, along with state-of-the-art calculations, can be used to determine  $r_p$ . Although the e-p results can be somewhat less precise than the spectroscopy results, until 2010 the values of  $r_{\rm D}$  obtained from these two methods<sup>2,5</sup> mostly agreed with each other<sup>10</sup>. Since that year, two new results based on Lamb shift measurements in  $\mu$ H were reported<sup>1,7</sup>. The Lamb shift in  $\mu$ H is several million times more sensitive to  $r_{\rm D}$  because the muon in a  $\mu H$  atom is about 200 times closer to the proton than is the electron in a H atom. To the surprise of both the nuclear and atomic physics communities, the two µH results<sup>1,7</sup>, displaying unprecedented precision with an estimated uncertainty of

Duke University and Triangle Universities Nuclear Laboratory, Durham, NC, USA. 2North Carolina A&T State University, Greensboro, NC, USA. 3Mississippi State University, Mississippi State, MS, USA. 4Idaho State University, Pocatello, ID, USA. 5University of Virginia, Charlottesville, VA, USA. 6Thomas Jefferson National Accelerator Facility, Newport News, VA, USA. 7Argonne National Laboratory, Lemont, IL, USA. <sup>8</sup>University of North Carolina, Wilmington, NC, USA. <sup>9</sup>Kharkov Institute of Physics and Technology, Kharkov, Ukraine. <sup>10</sup>Massachusetts Institute of Technology, Cambridge, MA, USA, 11Old Dominion University, Norfolk, VA, USA, 12Alikhanov Institute for Theoretical and Experimental Physics NRC "Kurchatov Institute", Moscow, Russia, 13University of Massachusetts, Amherst, MA, USA. 14 Hampton University, Hampton, VA, USA. 15 College of William and Mary, Williamsburg, VA, USA. 16 Norfolk State University, Norfolk, VA, USA. 17 Yerevan Physics Institute, Yerevan, Armenia, \*e-mail; gasparan@ilab.org; d.dutta@msstate.edu

#### **Article**

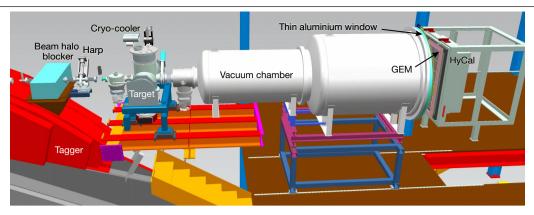


Fig. 1| The PRad experimental setup. A schematic layout of the PRad experimental setup in Hall B at Jefferson Laboratory, with the electron beam incident from the left. The key beam-line elements are shown along with the

windowless hydrogen gas target, the two-segment vacuum chamber and the two detector systems (see the Methods for a brief overview and the Supplementary Information for a description of the target and individual detectors).

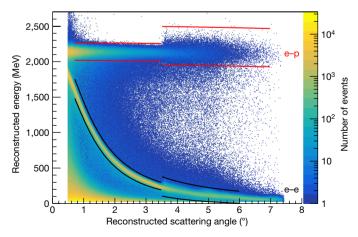
<0.1%, combined to be eight standard deviations smaller than the average value obtained from all previous experiments. This became known as the proton radius puzzle  $^{11}$ , unleashing intensive experimental and theoretical efforts aimed at resolving the disagreement.

The discrepancy between the values of  $r_{\rm p}$  as measured in H and  $\mu$ H atoms remains unresolved. Moreover, the two most recent H spectroscopy measurements disagree with each other  $^{3,4}$ , which has added a new dimension to and renewed the urgency of this problem. A fundamental difference between the e-p and  $\mu$ -p interactions could be the origin of the discrepancy; however, there are abundant experimental constraints on any such 'new physics', although models that resolve the puzzle by invoking new force carriers have been proposed  $^{11,12}$ . More mundane solutions continue to be explored: for example, it has been rigorously shown that the definition of  $r_{\rm p}$  used in all three major experimental approaches was consistent  $^{13}$ . The effect of two-photon exchange on  $\mu$ H spectroscopy  $^{14,15}$ , and form-factor nonlinearities in e-p scattering  $^{16-18}$  have also been examined. None of these studies has adequately explained the puzzle, reinforcing the need for additional high-precision measurements of  $r_{\rm p}$  that use new experimental techniques and different systematics.

The PRad collaboration at Jefferson Laboratory has developed and performed an e-p experiment as an independent measurement of  $r_{\rm p}$ to address the puzzle. The PRad experiment, in contrast with previous e-p experiments, was designed to use a magnetic-spectrometer-free. calorimeter-based method<sup>19</sup>. The design of the PRad experiment implemented three major improvements over previous e-p experiments. First, the large angular acceptance (0.7°-7.0°) of the hybrid calorimeter (HyCal) enabled large  $Q^2$  coverage, spanning two orders of magnitude  $(2.1 \times 10^{-4} \,\mathrm{GeV^2/c^2})$  to  $6 \times 10^{-2} \,\mathrm{GeV^2/c^2}$ , where c is the speed of light in a vacuum) in the low  $Q^2$  range. The fixed location of HyCal eliminated the many normalization parameters that plague magnetic-spectrometer-based experiments in which the spectrometer must be physically moved to many different angles to cover the desired range of  $Q^2$ . In addition, the PRad experiment reached extreme forward-scattering angles of down to  $0.7^{\circ}$ , achieving a  $Q^2$  value of  $2.1 \times 10^{-4}$  GeV<sup>2</sup>/ $c^2$ ; this is, to our knowledge, the lowest  $Q^2$  obtained from e-p experiments and is an order of magnitude lower than that previously achieved<sup>5</sup>. Reaching a lower range for  $Q^2$  is critical because  $r_D$  is determined from the slope of the electric form factor at  $Q^2 = 0$ . Second, the extracted e-p crosssections were normalized to the well-known quantum electrodynamics process  $e^-e^- \rightarrow e^-e^-$  (Møller scattering from atomic electrons,  $e^-e$ ), which was measured simultaneously alongside e-p scattering, using the same detector acceptance. This led to a substantial reduction in the systematic uncertainties of measuring the e-p cross-sections. Third, the background generated from the target windows, one of the dominant sources of systematic uncertainty in all previous e-p experiments, was highly suppressed in the PRad experiment.

The PRad experimental apparatus consisted of four main elements (Fig. 1). (1) A 4-cm-long windowless cryo-cooled hydrogen gas flow target with an areal density of  $2 \times 10^{18}$  atoms per cm², which eliminated the beam background from the target windows. (2) The high-resolution, large-acceptance hybrid electromagnetic calorimeter, HyCal²²². The complete azimuthal coverage of HyCal for the forward-scattering angles enabled simultaneous detection of the pair of electrons from e–e scattering. (3) A plane made of two high-resolution X-Y gas electron multiplier (GEM) coordinate detectors located in front of HyCal. (4) A two-section vacuum chamber spanning the 5.5-m distance from the target to the detectors.

The PRad experiment was performed in Hall B at Jefferson Laboratory in May–June of 2016, using 1.1-GeV and 2.2-GeV electron beams. The standard Hall B beam line, designed for low beam currents (0.1–50 nA), was used in this experiment. The incident electrons that scattered off the target protons and the Møller electron pairs were detected in the GEM detector and HyCal. The energy and position of the detected electron(s) were measured by HyCal, and the transverse (X-Y) position was measured by the GEM detector, which was used to assign the  $Q^2$  for each detected event. The GEM detector, which has a position resolution of 72  $\mu$ m, improved the measurement accuracy of  $Q^2$  compared to detection by HyCal alone. Furthermore, the GEM detector suppressed the contamination from photons generated in the target and other beam-line materials; HyCal is equally sensitive to electrons



**Fig. 2**| **Event reconstruction.** The reconstructed energy versus angle for e−p and e−e events for an electron beam energy of 2.2 GeV. The red and black lines indicate the event selections for e−p and e−e, respectively. The angles  $\leq$  3.5° are covered by the crystal PbWO $_4$  modules of HyCal and the larger angles by the Pb glass modules. The colour bar shows the number of events.

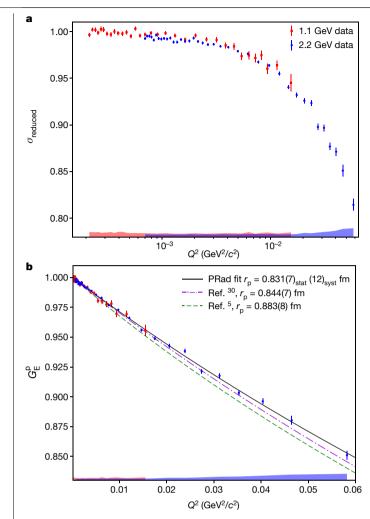


Fig. 3 | The measured cross-section and form factor. a, The reduced crosssection,  $\sigma_{\text{reduced}} = \left(\frac{d\sigma}{d\Omega}\right)_{e-p} / \left(\frac{d\sigma}{d\Omega}\right)_{\text{point-like}} ((4M_p^2 E'/E)/(4M_p^2 + Q^2))$ , (where E is the

electron beam energy, E' is the energy of the scattered electron,  $M_{\rm p}$  is the mass of the proton and  $\Omega$  is the solid angle subtended by the scattered electron detector), for the PRade-p data. Dividing out the kinematic factor inside the square brackets,  $\sigma_{\text{reduced}}$  is a linear combination of the electromagnetic form factors squared. The bands at the bottom of the plot are the size of the systematic uncertainties, for 1.1 GeV (red) and 2.2 GeV (blue). The error bars show statistical uncertainties. **b**,  $G_F^p$  as a function of  $Q^2$ . The data points are normalized by the parameter n in equation (1) for the 1.1-GeV and 2.2-GeV data, labelled as  $n_1$  and  $n_2$ , respectively. The error bars show statistical uncertainties. The bands are the systematic uncertainties as in a. The solid black curve shows  $G_{\rm F}^{\rm p}(Q^2)$  as a fit to the function given by equation (1). Also shown is the fit from a previous e-p experiment<sup>5</sup>, giving  $r_p = 0.883(8)$  fm (green dashed line) and another previous calculation  $^{30}$  giving  $r_D = 0.844(7)$  fm (purple dot-dashed line).

and photons, whereas the GEM detector is mostly insensitive to neutral particles. The GEM detector also helped to suppress position-dependent irregularities in the response of HyCal. A plot of the reconstructed energy versus the reconstructed angle for e-p and e-e events is shown in Fig. 2 for the 2.2-GeV beam energy.

The background was measured periodically with an empty target cell. To mimic the residual gas in the beam line,  $H_2$  gas at very low pressure was allowed in the target chamber during the empty target runs. The charge-normalized e-p and Møller scattering yields from the empty target cell were used to subtract the background contributions. The beam current was measured with the Hall B Faraday cup with an uncertainty of <0.1%<sup>21</sup>. Further details on the background subtraction can be found in the Supplementary Information.

A comprehensive Monte Carlo simulation of the PRad setup was developed using the Geant4 toolkit<sup>22</sup>. The simulation consists of two separate event generators built for the e-p and e-e processes<sup>23,24</sup>. Inelastic e-p scattering background events were also included in the simulation using a fit<sup>25</sup> to the e-p inelastic world data. The simulation included signal digitization and photon propagation, which were critical for the precise reconstruction of the position and energy of each event in the HyCal. The details are described in the Supplementary Information.

The e-p cross-sections were obtained by comparing the simulated and measured e-p yield relative to the simulated and measured e-e yield (see Supplementary Information for details). The extracted reduced cross-section is shown in Fig. 3a. The e-p elastic cross-section is related to  $G_{\rm E}^{\rm p}$  and the proton magnetic form factor,  $G_{\rm M}^{\rm p}$  by the Rosenbluth formula<sup>19</sup>. In the very low  $O^2$  region covered by the PRad experiment, the cross-section is dominated by the contribution from  $G_{\rm F}^{\rm p}$ . Thus, the uncertainty introduced from  $G_M^p$  is negligible. In fact, when using a wide variety of parametrizations  $^{5,26-28}$  for  $G_M^p$ , the extracted  $G_F^p$ varies by about 0.2% at  $Q^2 = 0.06 \,\text{GeV}^2/c^2$ , the largest  $Q^2$  accessed by the PRad experiment, and by <0.01% in the  $Q^2$  < 0.01 GeV<sup>2</sup>/ $c^2$  region. The largest variation in  $r_p$  arising from the choice of  $G_M^p$  parametrization is 0.001 fm.  $G_F^p(Q^2)$  as extracted from our data is shown in Fig. 3b, using the Kelly parametrization<sup>26</sup> for  $G_{\rm M}^{\rm p}$ .

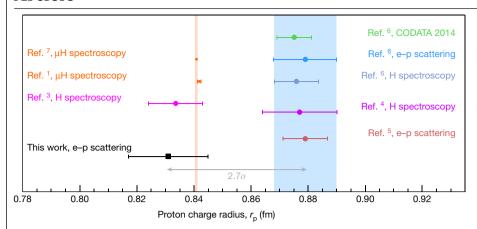
The slope of  $G_F^p(Q^2)$  as  $Q^2 \rightarrow 0$  is proportional to  $r_p^2$ . A common practice is to fit  $G_{\rm E}^{\rm p}(Q^2)$  to a functional form and to obtain  $\dot{r}_{\rm p}$  by extrapolating to  $Q^2 = 0$ . However, each functional form truncates the higher-order moments of  $G_{\rm F}^{\rm p}(Q^2)$  differently and introduces a model dependence that can bias the determination of  $r_p$ . It is critical to choose a robust functional form that is most likely to yield an unbiased estimation of  $r_{\rm p}$  given the uncertainties in the data, and to test the chosen functional form over a broad range of parametizations<sup>29</sup> of  $G_F^p(Q^2)$ . To simultaneously minimize possible bias in the determination of the radius and the total uncertainty, various functional forms were examined for their robustness in reproducing an input  $r_p$  used to generate a mock dataset with the same statistical uncertainty as the PRad data. The robustness, quantified as the root-mean square error (RMSE), is defined as RMSE =  $\sqrt{(\delta R)^2 + \sigma^2}$ , where  $\delta R$  is the bias or the difference between the input and extracted radius and  $\sigma$  is the statistical variation of the fit to the mock data<sup>29</sup>. Previous studies<sup>29</sup> show (see Supplementary Information) that consistent results with the smallest uncertainties can be achieved using a multi-parameter rational function, which we refer to as Rational(1,1):

$$f(Q^2) = nG_{\rm E}^{\rm p}(Q^2) = n\frac{1 + p_1 Q^2}{1 + p_2 Q^2} \tag{1}$$

where n is the floating normalization parameter,  $p_1$  and  $p_2$  are fit parameters and the proton charge radius is given by  $r_n = \sqrt{6(p_2 - p_1)}$ . The  $G_E^p(Q^2)$ , extracted from the 1.1-GeV and 2.2-GeV data, was fitted simultaneously using the Rational (1, 1) function. Independent normalization parameters  $n_1$  and  $n_2$  were assigned for the 1.1-GeV and 2.2-GeV data, respectively, to allow for differences in normalization uncertainties, but the  $Q^2$  dependence was identical. The parameters obtained from fits to the Rational (1,1) function are  $n_1 = 1.0002 \pm 0.0002_{\text{stat}} \pm 0.0020_{\text{syst}}$ ;  $n_2 = 0.9983 \pm 0.0002_{\text{stat}}$  $\pm 0.0013_{\text{syst}}$ ; and  $r_p = 0.831 \pm 0.007_{\text{stat}} \pm 0.012_{\text{syst}}$  fm. The Rational(1,1) function describes the data very well, with a reduced  $\chi^2$  of 1.3 when considering only the statistical uncertainty. The values of  $r_p$  for a variety of functional forms fitted to the PRad data are shown in Supplementary Fig. 15.

To determine the systematic uncertainty in  $r_p$ , a Monte Carlo technique was used to randomly smear the cross-section and  $G_F^p(Q^2)$  data points for each known source of systematic uncertainty. The value of  $r_{\rm p}$  was extracted from the smeared data and the process was repeated 100,000 times. The root-mean square of the resulting distribution of  $r_p$  is recorded as the systematic uncertainty. The dominant systematic uncertainties of  $r_{\rm p}$  are those that are  $Q^2$ -dependent, which primarily affect the lowest  $Q^2$  data: the Møller radiative corrections, the background subtraction for the

### **Article**



**Fig. 4** | **The proton charge radius.**  $r_p$  as extracted from the PRad data in this work, shown alongside other measurements of  $r_p$  since 2010 and previous CODATA recommended values. Our result is  $2.7\sigma$  smaller than the CODATA recommended value for e-p experiments  $^6$ . The orange and blue vertical bands show the uncertainty bounds of the  $\mu H$  and CODATA values for e-p scattering, respectively.

1.1-GeV data and event selection. The uncertainty in  $r_p$  arising from the finite  $Q^2$  range and the extrapolation to  $Q^2$  = 0 was investigated by varying the  $Q^2$  range of the mock dataset as part of the robustness study of the Rational(1,1) function<sup>29</sup>. This uncertainty was found to be much smaller than the relative statistical uncertainty, 0.8%. The total systematic relative uncertainty on  $r_p$  was found to be 1.4%, and is detailed in Supplementary Table 1 and described in the Supplementary Information.

The value of  $r_p$  obtained using the Rational(1, 1) function is shown in Fig. 4, with statistical and systematic uncertainties summed in quadrature. Our result, obtained from  $Q^2$  down to an unprecedented  $2.1 \times 10^{-4} \, \mathrm{GeV}^2/c^2$ , is about three standard deviations smaller than the previous high-precision electron scattering measurement<sup>5</sup>, which was limited to higher  $Q^2$  (>0.004  $\mathrm{GeV}^2/c^2$ ). However, our result is consistent with the  $\mu\mathrm{HLamb}$ -shift measurements<sup>1,7</sup>, and also with the recent 2S–4P transition-frequency measurement using ordinary H atoms<sup>3</sup>. Given that the lowest  $Q^2$  reached in the PRad experiment is an order of magnitude lower than in previous e–p experiments, and owing to the careful control of systematic effects, our result indicates that the proton radius is smaller than its previously accepted value from e–p measurements. Our result does not support any fundamental difference between e–p and  $\mu$ –p interactions and is consistent with the updated value announced for the Rydberg constant by CODATA<sup>8</sup>.

The PRad e-p experiment covers  $Q^2$  over two orders of magnitude in one setting. The experiment also exploited the simultaneous detection of e-p and e-e scattering to achieve good control of systematic uncertainties, which were, by design, different from previous e-p experiments. The extraction of  $r_{\rm p}$  using functional forms with validated robustness is another strength of this result. Our result demonstrates a large discrepancy with contemporary, high-precision e-p experiments. The result also implies that there is consistency between proton charge radii as obtained from e-p scattering measurements on ordinary hydrogen and spectroscopy of muonic hydrogen<sup>1,7</sup>. The PRad experiment demonstrates the clear advantages of the calorimeterbased method for determining  $r_{\rm p}$  from e-p experiments and points to further possible improvements in the accuracy of this method. It is also consistent with the recently announced shift in the Rydberg constant8, which has profound consequences, given that the Rydberg constant is one of the most precisely known constants of physics.

#### **Online content**

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1721-2.

- Pohl, R. et al. The size of the proton. Nature 466, 213–216 (2010).
- Mohr, P. J., Taylor, B. N. & Newell, D. B. CODATA recommended values of the fundamental physical constants: 2006. Rev. Mod. Phys. 80, 633-730 (2008).

- Beyer, A. et al. The Rydberg constant and proton size from atomic hydrogen. Science 358, 79–85 (2017).
- Fleurbaey, H. New measurement of the 1S-3S transition frequency of hydrogen: contribution to the proton charge radius puzzle. Phys. Rev. Lett. 120, 183001 (2018).
- Bernauer, J. C. et al. High-precision determination of the electric and magnetic form factors of the proton. *Phys. Rev. Lett.* 105, 242001 (2010).
- Mohr, P. J., Newell, D. B. & Taylor, B. N. CODATA recommended values of the fundamental physical constants: 2014. J. Phys. Chem. Ref. Data 45, 043102 (2016).
- Antognini, A. et al. Proton structure from the measurement of 2S-2P transition frequencies of muonic hydrogen. Science 339, 417-420 (2013).
- Mohr, P. J., Newell, D. B. & Taylor, B. N. CODATA recommended values of the fundamental physical constants: 2018. http://physics.nist.gov/constants (2019).
- Hasan, N. et al. Computing the nucleon charge and axial radii directly at Q<sup>2</sup> = 0 in lattice QCD. Phys. Rev. D 97, 034504 (2018).
- Mohr, P. J., Taylor, B. N. & Newell, D. B. CODATA recommended values of the fundamental physical constants: 2010. Rev. Mod. Phys. 84, 1527–1605 (2012).
- 11. Carlson, C. E. The proton radius puzzle. Prog. Part. Nucl. Phys. 82, 59-77 (2015).
- Liu, Y. S. & Miller, G. A. Validity of the Weizsäcker-Williams approximation and the analysis
  of beam dump experiments: production of an axion, a dark photon, or a new axial-vector
  boson. Phys. Rev. D 96, 016004 (2017).
- 13. Miller, G. A. Defining the proton radius: a unified treatment. Phys. Rev. C 99, 035202 (2019).
- Miller, G. A. Proton polarizability contribution: muonic hydrogen Lamb shift and elastic scattering. Phys. Lett. B 718, 1078–1082 (2013).
- Antognini, A. et al. Theory of the 2S-2P Lamb shift and 2S hyperfine splitting in muonic hydrogen. Ann. Phys. 331, 127-145 (2013).
- Lee, G., Arrington, J. R. & Hill, R. J. Extraction of the proton radius from electron–proton scattering data. Phys. Rev. D 92, 013013 (2015).
- Higinbotham, D. W. et al. Proton radius from electron scattering data. Phys. Rev. C 93, 055207 (2016).
- Griffioen, K., Carlson, C. & Maddox, S. Consistency of electron scattering data with a small proton radius. Phys. Rev. C 93, 065207 (2016).
- Gasparian, A. et al. High Precision Measurement of the Proton Charge Radius. Proposal to Jefferson Lab, PAC-38 C12-11-106 https://www.jlab.org/exp\_prog/proposals/11/PR12-11-106.pdf (2011).
- Gasparian, A. A high performance hybrid electromagnetic calorimeter at Jefferson Lab. In Proc. 11th Int. Conf. on Calorimetry in Particle Physics (eds Cecchi, C. et al.) 109–115 (World Scientific. 2005).
- Mecking, B. et al. The CEBAF large acceptance spectrometer (CLAS). Nucl. Instrum. Meth. A 503, 513–553 (2003).
- Agostinelli, S. et al. GEANT4: a simulation toolkit. Nucl. Instrum. Meth. A 506, 250–303 (2003).
- Akushevich, I., Gao, H., Ilyichev, A. & Meziane, M. Radiative corrections beyond the ultra relativistic limit in unpolarized ep elastic and Møller scatterings for the PRad experiment at Jefferson Laboratory. Eur. Phys. J. A 51, 1 (2015).
- Gramolin, A. V. et al. A new event generator for the elastic scattering of charged leptons on protons. J. Phys. G Nucl. Phys. 41, 115001 (2014).
- Christy, M. E. & Bosted, P. E. Empirical fit to precision inclusive electron–proton crosssections in the resonance region. *Phys. Rev. C* 81, 055213 (2010).
- 26. Kelly, J. J. Simple parametrization of nucleon form factors. Phys. Rev. C 70, 068202 (2004).
- Venkat, S., Arrington, J., Miller, G. A. & Zhan, X. Realistic transverse images of the proton charge and magnetic densities. *Phys. Rev. C* 83, 015203 (2011).
- Higinbotham, D. W. & McClellan, R. E. How analytic choices can affect the extraction of electromagnetic form factors from elastic electron scattering cross section data. Preprint at https://arxiv.org/abs/1902.08185 (2018).
- Yan, X. et al. Robust extraction of the proton charge radius from electron–proton scattering data. Phys. Rev. C 98, 025204 (2018).
- Alarcón, J. M., Higinbotham, D. W., Weiss, C. & Ye, Z. Proton charge radius from electron scattering data using dispersively improved chiral effective field theory. Phys. Rev. C 99, 044303 (2019).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© This is a U.S. government work and not under copyright protection in the U.S.; foreign copyright protection may apply 2019

#### Methods

The PRad experiment was conducted with 1.1-GeV and 2.2-GeV electron beams from the Continuous Electron Beam Accelerator Facility (CEBAF) accelerator incident on cold hydrogen atoms flowing through a windowless target cell. The scattered electrons, after traversing the vacuum chamber, were detected in the GEM detector and HyCal. They included electrons from elastic e-p scattering and e-e Møller scattering processes. The transverse (X-Y) positions measured by the GEM detector were used to calculate the  $Q^2$  value for each event. The e-p and e-e yields were obtained using appropriate cuts on the energy deposited in HyCal and the reconstructed angle. The e-p and e-e yields were binned as a function of  $Q^2$ . A comprehensive Monte Carlo simulation of the PRad experiment was used to extract the next-to-leading order e-p crosssection from the experimental yields. The e-p cross-sections were obtained by comparing the simulated and measured e-p yield relative to the simulated and measured Møller scattering yield. The value of  $G_{\rm F}^{\rm p}$ was extracted from the e-p cross-section using the Rosenbluth formula, and using a parametrization of  $G_{\rm M}^{\rm p}$ . The proton charge radius,  $r_{\rm p}$ , was obtained from the extracted  $G_F^p(Q^2)$  by fitting to the Rational (1, 1) functional form and extrapolating to  $Q^2 = 0$ . The Rational (1, 1) functional form was shown to be the most robust function for radius extraction from the PRad data, giving consistent results with the smallest uncertainties. See Supplementary Information for further details.

#### **Data availability**

The raw data from this experiment are archived in Jefferson Laboratory's mass storage silo.

#### **Code availability**

All computer codes used for data analysis and simulation are archived in lefferson Laboratory's mass storage silo.

Acknowledgements This work was funded in part by the US National Science Foundation (NSF MRI PHY-1229153) and by the US Department of Energy (contract number DE-FGO2-O3ER41231), including contract number DE-ACO5-O6OR23177, under which Jefferson Science Associates, LLC operates the Thomas Jefferson National Accelerator Facility. We thank the staff of Jefferson Laboratory for their support throughout the experiment. We are also grateful to all grant agencies for providing funding support to the authors throughout this project. We acknowledge discussions about radiative corrections with A. Afanasev, I. Akushevich, A. V. Gramolin and O. Tomalak. We thank S. Danagoulian for helping to restore the light monitoring system of HyCal. We also thank S. Javalkar for help with a beam halo study.

Author contributions A.G. is the spokesperson of the experiment. H.G., D. Dutta and M.K. are co-spokespersons of the experiment. A.G. developed the initial concepts of the experiment. A.G., H.G., D. Dutta and M.K designed and proposed the experiment. The entire PRad collaboration constructed the experiment and worked on the data collection. The COMSOL simulation of the target was built by Y.Z. The Monte Carlo simulation was built and validated by C. Peng, C.G., W.X. and X.B. with input from numerous other members of the collaboration. Calibrations were carried out by W.X., M.L., X.B., C. Peng, L.Y. and X.Y., with input from I.L. Analysis software tools were developed by C. Peng, with input from X.B., M.L., I.L., L.Y., W.X. and X.Y. The data analysis was carried out by W.X., C. Peng, X.B., M.L. and C.G., with input from A.G., H.G., D. Dutta, M.K., N.L., E.P., X.Y., D.W.H., L.Y. and M.L.K. All authors reviewed the manuscript.

Competing interests The authors declare no competing interests.

#### Additional information

**Supplementary information** is available for this paper at https://doi.org/10.1038/s41586-019-1701-2

Correspondence and requests for materials should be addressed to A.G. or D.D.

Peer review information Nature thanks Krzysztof Pachucki and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints.

# Carrier-resolved photo-Hall effect

https://doi.org/10.1038/s41586-019-1632-2

Received: 2 April 2018

Accepted: 1 August 2019

Published online: 7 October 2019

Oki Gunawan<sup>1\*</sup>, Seong Ryul Pae<sup>2</sup>, Douglas M. Bishop<sup>1</sup>, Yudistira Virgus<sup>1</sup>, Jun Hong Noh<sup>3,4</sup>, Nam Joong Jeon<sup>3</sup>, Yun Seog Lee<sup>1,5</sup>, Xiaoyan Shao<sup>1</sup>, Teodor Todorov<sup>1</sup>, David B. Mitzi<sup>6,7</sup> & Byungha Shin<sup>2</sup>\*

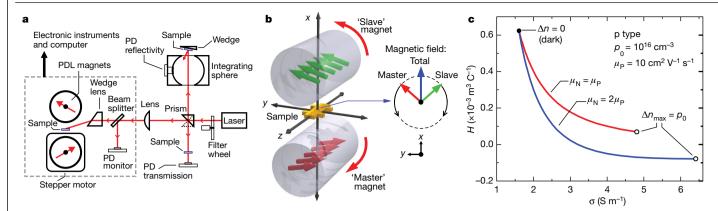
The fundamental parameters of majority and minority charge carriers—including their type, density and mobility–govern the performance of semiconductor devices yet can be difficult to measure. Although the Hall measurement technique is currently the standard for extracting the properties of majority carriers, those of minority carriers have typically only been accessible through the application of separate techniques. Here we demonstrate an extension to the classic Hall measurement—a carrier-resolved photo-Hall technique—that enables us to simultaneously obtain the mobility and concentration of both majority and minority carriers, as well as the recombination lifetime, diffusion length and recombination coefficient. This is enabled by advances in a.c.-field Hall measurement using a rotating parallel dipole line system and an equation,  $\Delta \mu_{\rm H} = {\rm d}(\sigma^2 H)/{\rm d}\sigma$ , which relates the hole–electron Hall mobility difference ( $\Delta \mu_H$ ), the conductivity ( $\sigma$ ) and the Hall coefficient (H). We apply this technique to various solar absorbers-including high-performance lead-iodidebased perovskites—and demonstrate simultaneous access to majority and minority carrier parameters and map the results against varying light intensities. This information, which is buried within the photo-Hall measurement<sup>1,2</sup>, had remained inaccessible since the original discovery of the Hall effect in  $1879^3$ . The simultaneous measurement of majority and minority carriers should have broad applications, including in photovoltaics and other optoelectronic devices.

The Hall effect measurement is one of the most important characterization techniques for electronic materials, and the effect has become the basis of fundamental advances in condensed matter physics, such as the integer and fractional quantum Hall effects<sup>4,5</sup>. The measurements reveal fundamental information about the majority charge carrier—that is, its type (p or n), density and mobility. In a solar cell, the parameters of the majority carrier determine the overall device architecture, the width of the depletion region and the bulk series resistance. The properties of the minority carrier, however, determine other key parameters that directly affect the overall performance of the device, such as recombination lifetime  $(\tau)$ , diffusion length  $(L_D)$  and recombination coefficients  $(k_n)$ . Unfortunately, the standard Hall measurement yields information regarding only the majority carrier. Attempts to measure the properties of both majority and minority carriers in high-performance light-absorbing materials have been made; however, they require a wide range of experimental techniques that typically use different sample configurations and illumination levels, thereby presenting additional complications in the analysis 6-14 (Supplementary Information sections F, G). The extraction of reliable information on charge carriers is particularly sought after in the study of organic-inorganic hybrid perovskites. This family of materials is currently receiving intense attention, owing to rapid progress in their application in high-performance solar cells-the current record power conversion efficiency (PCE) for devices containing such materials is 25.2% 15 – as well as in other optoelectronic devices, including light-emitting diodes<sup>16</sup> and photodetectors<sup>17</sup>. A full understanding of the charge-transport properties of perovskites will help to elucidate the operating principles of devices that contain these materials, thereby guiding their further improvement.

In this work we present a carrier-resolved photo-Hall (CRPH) measurement technique that is capable of simultaneously extracting the mobilities, densities and subsequent derivative parameters  $(\tau, L_D)$  of both majority and minority carriers as a function of light intensity. This technique relies on two key elements: an equation that yields the difference between the Hall mobilities of the hole and electron, and a high-sensitivity Hall measurement using a parallel dipole line (PDL) a.c. Hall system<sup>18</sup> (Fig. 1a, b). In the classic Hall measurement without illumination, three parameters can be obtained for majority carriers: the type (p or n), from the sign of Hall coefficient H; the carrier density  $(n_C = r/He)$ ; and the Hall mobility  $(\mu_H = \sigma H)$ ; where e is the electron charge and r is the Hall scattering factor. The key challenge in the photo-Hall transport problem-that is, extracting information from the majority and minority carriers—requires solving for three unknowns at a given illumination level: hole and electron (drift) mobility ( $\mu_P$ ,  $\mu_N$ ) and their photocarrier densities  $(\Delta n, \Delta p)$ , which are equal under steady-state conditions. Unfortunately, we have only two measured quantities:  $\sigma$ and H, as a function of illumination. The key insight into solving this problem is illustrated in Fig. 1c. We consider two p-type systems with the same majority carrier density  $(p_0)$  and mobility  $(\mu_P)$  but different minority carrier mobilities ( $\mu_N$ ). When these systems are excited with the same photocarrier density,  $\Delta n_{\text{max}}$ , they will produce different  $\sigma$ –H

1BM T. J. Watson Research Center, Yorktown Heights, NY, USA. 2Department of Materials Science and Engineering, Korea Advanced Institute of Science and Technology, Daejeon, South Korea. <sup>3</sup>Division of Advanced Materials, Korea Research Institute of Chemical Technology, Daejeon, South Korea, <sup>4</sup>School of Civil, Environmental and Architectural Engineering and Green School, Korea University, Seoul, South Korea. Department of Mechanical and Aerospace Engineering, Seoul National University, Seoul, South Korea. Department of Mechanical Engineering and Material Science, Duke University, Durham, NC, USA, 7Department of Chemistry, Duke University, Durham, NC, USA, \*e-mail: oqunawa@us.ibm.com; byungha@kaist.ac.kr

#### **Article**



**Fig. 1**| **The carrier-resolved photo-Hall measurement. a**, The PDL photo Hall setup for a complete photo-Hall experiment. **b**, The rotating PDL magnet system that generates a unidirectional and single harmonic a.c. magnetic field at the centre (see animation in Supplementary Video 1). **c**, Theoretical

calculation of two p-type systems with the same majority mobility ( $\mu_{\rm P}$ ) but different minority mobility ( $\mu_{\rm N}$ ) under increasing illumination, yielding different conductivity–Hall coefficient ( $\sigma$ –H) curves. The slope of the  $\sigma$ –H curve contains the information of  $\Delta\mu_{\rm H}$ .

curves owing to the increasing role of the minority carrier in the total conductivity, even though they start from the same point in the dark. Therefore, the characteristics of the  $\sigma$ -H curves—specifically the slope (dH/d $\sigma$ )—contains detailed information about the mobilities of the two systems. We show that the Hall mobility difference,  $\Delta \mu_{\rm H} = r \Delta \mu = r (\mu_{\rm P} - \mu_{\rm N})$ , is given as (Supplementary Information section B):

$$\Delta \mu_{\rm H} = \frac{\mathrm{d}(\sigma^2 H)}{\mathrm{d}\sigma} = \left(2 + \frac{\mathrm{d}\ln H}{\mathrm{d}\ln\sigma}\right)\sigma H \tag{1}$$

Note that  $\sigma$  and H are experimentally obtained as a function of varying light intensity or photocarrier density  $\Delta n$ ; however, fortuitously, the  $\Delta n$ term cancels out of equation (1). There are two equivalent expressions for  $\Delta\mu_{\rm H}$  in equation (1), which enable slope analysis for low and high injection. The term  $d \ln H/d \ln \sigma$  has special experimental meaning, as shown for the perovskite example discussed later. This equation applies to both p- and n-type materials and assumes that the dark carrier densities  $(p_0 \text{ or } n_0)$  are fixed, that  $\Delta n = \Delta p$  under steady-state conditions, and that the mobilities are constant as a function of light intensity (see Supplementary Information section B.2 for a generalized model in which mobilities vary with illumination). The Hall scattering factor r generally lies between 1 and 2. It approaches 1 at high magnetic field and generally is assumed<sup>19</sup> to be 1, including in this work. Using the known two-carrier expressions in the low-magnetic-field regime<sup>20</sup>  $(B \ll 1/\mu)$ —that is,  $\sigma = e(p\mu_P + n\mu_N)$  and  $H = r(p - \beta^2 n)/(p + \beta n)^2 e$ , where p and n are hole and electron densities and  $\beta = \mu_N/\mu_P$  is the mobility ratio—we can completely solve the photo-Hall transport problem, for example, for a p-type material:

$$\beta = \frac{2\sigma(r\Delta\mu - \sigma H) - re\Delta\mu^2 p_0 \pm \Delta\mu\sqrt{rep_0}\sqrt{re\Delta\mu^2 p_0 + 4\sigma(\sigma H - r\Delta\mu)}}{2\sigma(r\Delta\mu - \sigma H)} \quad (2)$$

$$\Delta n = \frac{\sigma(1-\beta) - e\Delta\mu p_0}{e\Delta\mu(1+\beta)} \tag{3}$$

Finally, we obtain  $\mu_P = \Delta \mu/(1-\beta)$  and  $\mu_N = \beta \mu_P$ . Note that we need to know the background hole density,  $p_0$ , from the dark measurement. Equations (1)–(3) are referred to as the  $\Delta \mu$  calculation model.

The second requirement to enable the CRPH measurement involves obtaining a clean Hall signal. Unfortunately, in many photovoltaic films, high sample resistance ( $R > 10~\text{G}\Omega$ )—as in the case of perovskites—or low mobility ( $\mu < 1~\text{cm}^2~\text{V}^{-1}~\text{s}^{-1}$ ) can produce noisy Hall signals. Therefore, a.c.-field Hall techniques coupled with Fourier analysis and lock-in detection are crucial. We recently developed a high sensitivity a.c.-field

Hall system based on a rotating PDL magnet system  $^{18,21,22}$ . The PDL system is a recently discovered natural magnetic trap that harbours a field-confinement effect that generates a magnetic camelback potential along its longitudinal axis $^{23}$ . This effect is used to optimize the field uniformity (Supplementary Information section A). The PDL Hall system consists of a pair of diametric cylindrical magnets separated by a gap. One magnet (the 'master') is driven by a motor and another (the 'slave') follows in the opposite direction. This system produces a unidirectional and single harmonic field at the centre, where the sample resides (Fig. 1b, Supplementary Video 1), which forms the basis for a successful photo-Hall experiment. As well as the photo-Hall measurement, optical measurements (for example, transmission and reflectivity) to calculate the absorbed photon density ( $G_{\gamma}$ ) can also be performed using the same setup (Fig. 1a, Supplementary Information section C).

To demonstrate the CPRH technique, two examples are discussed in detail: a lead-halide-based perovskite film and a silicon sample, which serve as high  $(\Delta n > p_0)$  and low  $(\Delta n < p_0)$  injection cases, respectively. The first example uses a (FA,MA)Pb(I,Br)<sub>3</sub> (FA, formamidinium; MA, methylammonium) perovskite film, which was fabricated using the same method that produced a recent record PCE<sup>24</sup>, but with further optimization of the process. A companion device in the same batch vielded a PCE of 20.8% (Methods). The measurement device is a six-terminal Hall bar with an active area of 2 mm × 4 mm and a film thickness d of 0.55  $\mu$ m as shown in Fig. 2b. First, we measured the sample in the dark and obtained the properties of the majority carrier: p type,  $p_0 = 8.3 \times 10^{11} \,\text{cm}^{-3}$  and  $\mu_P = 9.8 \,\text{cm}^2 \,\text{V}^{-1} \,\text{s}^{-1}$ . Next, we performed the measurements under several laser intensities (up to about 40 mW cm<sup>-2</sup>; wavelength  $\lambda = 638$  nm). Examples of longitudinal ( $R_{xx}$ ) and transverse  $(R_{xy})$  magnetoresistance under light are shown in Fig. 2a. The  $R_{xy}$  trace shows the expected Hall signal with a Fourier component at the same frequency  $(f_{ref})$  as the magnetic field B (Fig. 2a). The desired Hall signal  $R_{\rm H}$  is obtained using numerical lock-in detection <sup>18</sup> based on a reference sinusoidal signal with the same phase as B (Fig. 2a). The  $\sigma$  and H values are then calculated from  $R_{xx}$  and  $R_H$ . We also observe a second harmonic component at  $2f_{ref}$  in the  $R_{xy}$  Fourier spectrum, which is also evident in the original  $R_{xy}$  trace as a double frequency oscillation (Fig. 2a). This component is not the desired Hall signal and thus is rejected. It arises from another magnetoresistance effect<sup>25</sup>, which is stronger in  $R_{xx}$ , and also appears in  $R_{xy}$  because of  $R_{xx}$ - $R_{xy}$  mixing due to the finite size of the Hall bar contact arms. This highlights the importance of inspecting the Fourier spectrum of the Hall signal and of using lock-in detection, as opposed to simple amplitude measurement.

The measurement returns a series of  $\sigma$  and H points that change substantially upon illumination (Fig. 2b):  $\sigma$  increases by a factor of around 340 and H decreases by a factor of around 1,400. Our photo-Hall

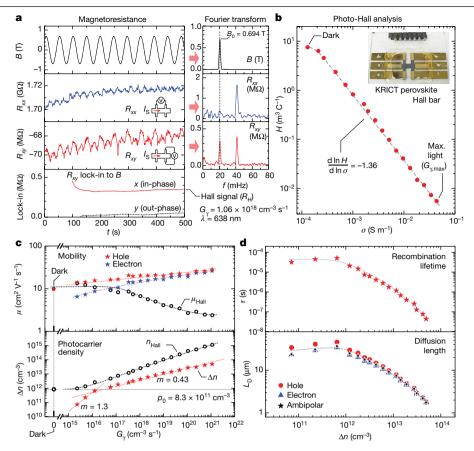


Fig. 2 | Carrier-resolved photo-Hall analysis in a high-performance **perovskite film. a**, Magnetoresistance sweep ( $R_{xx}$ , longitudinal;  $R_{xy}$ , transverse), Fourier transform and lock-in detection of the Hall signal  $(R_{xy})$ .  $\mathbf{b}$ ,  $\sigma$ -H plot for photo-Hall analysis. Inset, the perovskite Hall bar device. **c**, Majority  $(\mu_P)$  and minority  $(\mu_N)$  carrier mobility and photocarrier density  $\Delta n$ 

plotted against absorbed photon density  $G_{v}$ , with  $n_{\text{Hall}}$  and  $\mu_{\text{Hall}}$  denoting the single-carrier Hall density and mobility. The background carrier density  $p_0$  is indicated by a grey line. **d**, Recombination lifetime ( $\tau$ ) and diffusion length ( $L_D$ ) mapped against  $\Delta n$ . All dashed curves are guides for the eye.

equation (equation (1)) provides a simple and quick insight into the data by looking at the slope of the  $\sigma$ -H data using the log scale. If the slope  $(d \ln H/d \ln \sigma)$  is equal to -2, then  $\mu_P = \mu_{Nr}$ , if it is larger (less) than -2while H is positive, then  $\mu_P > \mu_N (\mu_P < \mu_N)$ . From Fig. 2b we obtain the overall d ln H/d ln  $\sigma$  = -1.36, which implies that  $\mu_P > \mu_N$ . Furthermore, we can evaluate  $\Delta \mu_{\rm H}$  at any  $\sigma$ -H point—for example, at the maximum light intensity:  $\Delta \mu_H = 1.9 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ . We proceed to solve for  $\mu_P$ ,  $\mu_N$  and  $\Delta n$  using the previously discussed  $\Delta\mu$  model and plot the values with respect to  $G_v$  (Fig. 2c). Owing to a considerable change in mobility with light intensity in the perovskite sample, we used the generalized  $\Delta\mu$  model (as discussed in Supplementary Information section B.2). Given the varying mobilities, this model introduces a correction that yields final mobility values as much as two times smaller than the initial mobility values at the highest light intensity. We obtain the final solution  $\mu_P = (14-28)$ cm<sup>2</sup> V<sup>-1</sup> s<sup>-1</sup> and  $\mu_N = (7-26)$  cm<sup>2</sup> V<sup>-1</sup> s<sup>-1</sup>; both  $\mu_P$  and  $\mu_N$  increase with  $G_V$ . This increase could arise from a light-modulated intragranular barrier effect<sup>1</sup>. We further obtained  $\Delta n$ , which increases with  $G_{v}$  as expected. For comparison, we also plot the 'single-carrier' Hall mobility ( $\mu_H = \sigma H$ ) and Hall density  $(n_H = 1/eH)$ , which have often been used to estimate  $\mu$  and  $\Delta n$  in previous photo-Hall studies<sup>13</sup>. As seen in Fig. 2c, these estimates can be very different from the actual  $\mu_P$ ,  $\mu_N$  and  $\Delta n$  values obtained from the CRPH measurement.

In addition to the basic properties of the majority and minority carriers, we then investigated the recombination mechanism in detail by plotting  $\Delta n$  against  $G_v$ , as shown in Fig. 2c. The data show two power-law regimes following  $\Delta n \approx G_v^m$  with  $m \approx 1$  and  $m \approx 0.5$ . The m = 1 (m = 0.5) behaviour is expected for a monomolecular recombination (bimolecular recombination) regime<sup>13</sup>; however, in this case the  $m \approx 0.5$  regime is more likely explained by trapping. Consider, for example, a single-level trap model. The lifetime—for example, for a hole—is given as:  $\tau = 1/C_P n_r$ , where  $C_{\rm P}$  represents the capture cross-section for a hole and  $n_{\rm r}$  is the density of the trapped electrons. At very low light intensities, the lifetime is constant because  $n_r$  is dominated by the equilibrium (dark) level of charged electron traps,  $n_r = n_{r0}$ . As the light intensity increases, the number of charged traps increases owing to the increase in injected electrons. This in turn reduces  $\tau$  and explains the low exponent ( $m \approx 0.5$ ) seen<sup>26</sup> in Fig. 2c. The alternative explanation of bimolecular recombination can be discarded because the maximum photocarrier density in our experiment ( $\Delta n \approx 10^{14} \,\mathrm{cm}^{-3}$ ) is around 1,000 times lower than the typical density required ( $\Delta n \approx 10^{17} \, \text{cm}^{-3}$ ) in order for the bimolecular recombination to dominate in perovskite<sup>27,28</sup> (Supplementary Information section D).

The measurement also provides access to the recombination lifetime,  $\tau = \Delta n/G$  and the carrier diffusion length,  $L_{\rm D} = \sqrt{k_{\rm B}T\mu\tau/e}$ , where  $k_{\rm B}$  is the Boltzmann constant and T is temperature. G is the photocarrier generation rate given as  $G = \eta G_{v}$ , where  $\eta$  is the photocarrier generation efficiency, which is often assumed to be unity. At high injection level when  $\Delta n$ ,  $\Delta p > p_0$ , it is more appropriate to use an ambipolar diffusion length<sup>29</sup>, which can also be calculated from our CRPH data:  $L_{\rm D,am} = \sqrt{k_B T \tau (n+p)/e(n/\mu_{\rm p} + p/\mu_{\rm N})}$ . Furthermore, we can plot these results as a function of  $\Delta n$  (Fig. 2d). The hole, electron and ambipolar diffusion lengths fall very close to each other given similar hole and electron mobilities. From this analysis, we obtain values of  $\tau$  of up to 40  $\mu$ s and  $L_D$  values of around 30  $\mu$ m at the lowest light intensity. However, these values vary markedly, and decrease to 44 ns and 1.7 μm, respectively, at the highest light intensity. The relatively high values of  $\tau$  and  $L_D$  obtained in this study attest to the high quality of this

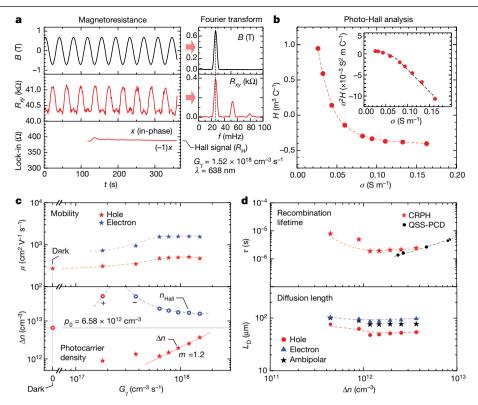


Fig. 3 | Carrier-resolved photo-Hall analysis in a single-crystal p-silicon sample. a, Transverse magnetoresistance sweep ( $R_{xy}$ ), Fourier transform and lock-in detection of the Hall signal. b,  $\sigma$ -H plot for photo-Hall analysis. The inset shows the equivalent plot in the form of  $\sigma^2H$  against  $\sigma$ . c, Majority ( $\mu_P$ ) and minority ( $\mu_N$ ) mobility and photocarrier density  $\Delta n$  plotted against absorbed

photon density  $G_{\gamma}$ .  $n_{\text{Hall}}$  is the single-carrier Hall density.  $\mathbf{d}$ , Recombination lifetime and minority carrier diffusion length plotted against  $\Delta n$ .  $\tau$  values measured using the quasi-steady-state photoconductance decay (QSS-PCD) technique are shown as black circles. All dashed curves are guides for the eye.

perovskite film<sup>24</sup>. We also compare our results with recent transport studies for perovskites (Supplementary Table 3) and obtain general agreement. We highlight that, given the large variation in  $\tau$  and  $L_{\rm D}$  with  $G_{\rm Y}$ , it is crucial to state the values of  $G_{\rm Y}$  or  $\Delta n$  when reporting these measurements.

In the second example, we investigate a single-crystal silicon sample. The sample is a Hall bar made of B-doped, Czochralski-grown silicon with active area of 3 mm × 3 mm and a thickness d of 725  $\mu$ m. This study demonstrates CRPH measurement of a well-known material, in the low injection regime and with a large thickness ( $d \gg 1/\alpha$  and  $L_{\rm D}$ ). We used a laser with a wavelength of 638 nm and an intensity of up to 50 mW cm<sup>-2</sup>. First, we obtain the  $\sigma$ -H curve that begins with a positive H value in the dark, indicating a p-type material with  $p_0$  = 6.6 × 10<sup>12</sup> cm<sup>-3</sup> (Fig. 3b). At higher light intensity, H becomes negative, indicating increasing electron (minority) carrier conductivity. For convenient extraction of  $\Delta\mu_{\rm H}$  we plot  $\sigma^2H$  against  $\sigma$  (Fig. 3b, inset), which is more appropriate for low-injection analysis. The data shows a monotonic behaviour with nearly constant slope at a high-intensity regime, which yields  $\Delta\mu_{\rm H}$  = -1,070 cm<sup>2</sup> V<sup>-1</sup> s<sup>-1</sup>.

We then calculated  $\mu_P$ ,  $\mu_N$  and  $\Delta n$  using equations (2) and (3), with the results shown in Fig. 3c, d. We obtain an average majority mobility of  $\mu_P = 486 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ , and a minority mobility of  $\mu_N = 1,560 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ ; these values are in good agreement with the hole (around  $500 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ ) and electron (around  $1,500 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ ) mobilities in silicon<sup>20</sup>. These values are sufficiently constant as a function of light intensity that we do not need to attempt the mobility-variation correction using the generalized model as for the perovskite analysis. We also plotted  $\Delta n$  against  $G_\gamma$  and obtained a curve that follows  $\Delta n \approx G_\gamma^m$  with m=1,2; this suggests a monomolecular recombination regime, as expected for silicon<sup>20</sup>. At the highest light intensity, we obtain  $\tau \approx 2 \mu \text{s}$  and  $L_{D,N} \approx 90 \mu \text{m}$ . For comparison, we also measured the lifetime using a quasi-steady-state photoconductance technique<sup>30</sup>; this yields  $\tau$  values of  $1-5 \mu \text{s}$  for

 $\Delta n = 2 \times 10^{12} - 10^{13}$  cm<sup>-3</sup>, in close agreement with the CRPH result (Fig. 3d, Methods). As an additional example of CRPH measurement in the low-injection regime we studied the material kesterite (Cu<sub>2</sub>ZnSn(S,Se)<sub>4</sub>), which is also of high interest for photovoltaics applications (Supplementary Information sections E, G).

In contrast to the classic Hall effect, which only yields three parameters, the CRPH technique yields 7N parameters:  $\mu_P$ ,  $\mu_N$ ,  $\Delta n$ ,  $\tau$ ,  $L_{D,N}$ ,  $L_{D,P}$  and  $L_{D,am}$  repeated at N light intensities. Additionally, it is also possible to calculate the relevant recombination coefficient; for example,  $k_1$ =1/ $\tau$  in the monomolecular recombination regime. Of the many electrical transport measurements performed on perovskites (as summarized in Supplementary Table 3), this is the first time to our knowledge that all minority and majority carrier characteristics have been determined simultaneously from a single experimental setup, on a single sample and mapped against varying light intensities under steady-state conditions. This demonstrates the power of the CRPH technique and represents a considerable expansion of the original Hall effect measurement<sup>3</sup>. The approach should also provide a valuable means of investigating the charge-carrier parameters of a wide range of conventional and emerging semiconductors for solar cells and broader applications.

### **Online content**

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1632-2.

 Fowler, A. Photo-Hall effect in CdSe sintered photoconductors. J. Phys. Chem. Solids 22, 181–188 (1961).

- Dresner, J. The photo-Hall effect in vitreous selenium. J. Phys. Chem. Solids 25, 505-511
- Hall, E. H. On a new action of the magnet on electric currents. Am. J. Math. 2, 287-292 (1879).
- Klitzing, K. v., Dorda, G. & Pepper, M. New method for high-accuracy determination of the fine-structure constant based on quantized Hall resistance. Phys. Rev. Lett. 45, 494-497 (1980).
- Tsui, D. C., Stormer, H. L. & Gossard, A. C. Two-dimensional magnetotransport in the extreme quantum limit. Phys. Rev. Lett. 48, 1559-1562 (1982).
- Ponseca, C. S., Jr et al. Organometal halide perovskite solar cell materials rationalized: 6. ultrafast charge generation, high and microsecond-long balanced mobilities, and slow recombination. J. Am. Chem. Soc. 136, 5189-5192 (2014).
- Leijtens, T. et al. Electronic properties of meso-superstructured and planar organometal 7 halide perovskite films: charge trapping, photodoping, and carrier mobility. ACS Nano 8, 7147-7155 (2014).
- Xing, G. et al. Long-range balanced electron-and hole-transport lengths in organic-8. inorganic CH<sub>2</sub>NH<sub>2</sub>Pbl<sub>3</sub>. Science **342**, 344-347 (2013).
- 9. Stranks, S. D. et al. Electron-hole diffusion lengths exceeding 1 micrometer in an organometal trihalide perovskite absorber, Science 342, 341-344 (2013).
- 10. Wehrenfennig, C., Liu, M., Snaith, H. J., Johnston, M. B. & Herz, L. M. Charge-carrier  $dynamics\ in\ vapour-deposited\ films\ of\ the\ organolead\ halide\ perovskite\ CH_3NH_3PbI_{3-x}Cl_{x^*}$ Energy Environ. Sci. 7, 2269-2275 (2014).
- Shi, D. et al. Low trap-state density and long carrier diffusion in organolead trihalide 11. perovskite single crystals. Science 347, 519-522 (2015).
- Dong, Q. et al. Electron-hole diffusion lengths >175 µm in solution-grown CH<sub>3</sub>NH<sub>3</sub>PbI<sub>3</sub> single crystals. Science 347, 967-970 (2015).
- 13. Chen, Y. et al. Extended carrier lifetimes and diffusion in hybrid perovskites revealed by Hall effect and photoconductivity measurements. Nat. Commun. 7, 12253 (2016).
- Gokmen, T., Gunawan, O. & Mitzi, D. B. Minority carrier diffusion length extraction in Cu<sub>2</sub>ZnSn(Se,S)<sub>4</sub> solar cells. J. Appl. Phys. 114, 114511 (2013).
- $National\ Renewable\ Energy\ Laboratories.\ \textit{Best research-cell efficiencies}\ https://www.nrel.$ gov/pv/cell-efficiency.html (2019).
- Tan, Z.-K. et al. Bright light-emitting diodes based on organometal halide perovskite. Nat. Nanotech. 9, 687-692 (2014).
- Dou, L. et al. Solution-processed hybrid perovskite photodetectors with high detectivity. Nat. Commun. 5, 5404 (2014).

- Gunawan, O., Virgus, Y. & Tai, K. F. A parallel dipole line system. Appl. Phys. Lett. 106, 062407 (2015)
- 19. Sze, S. M. & Ng, K. K. Physics of Semiconductor Devices 3rd edn (John Wiley & Sons, 2006).
- Schroder, D. K. Semiconductor Material and Device Characterization 3rd edn (John Wiley 20. & Sons. 2006).
- Gunawan, O. & Gokmen, T. Hall measurement system with rotary magnet. US patent 9.041,389 (2015).
- Gunawan, O. & Pereira, M. Rotating magnetic field Hall measurement system. US patent
- Gunawan, O. & Virgus, Y. The one-dimensional camelback potential in the parallel dipole line trap: Stability conditions and finite size effect. J. Appl. Phys. 121, 133902 (2017).
- Jeon, N. J. et al. Compositional engineering of perovskite materials for high-performance solar cells. Nature 517, 476-480 (2015).
- Zhang, C. et al. Magnetic field effects in hybrid perovskite devices. Nat. Phys. 11, 427-434 25. (2015).
- 26. Levine, I, et al. Can we use time-resolved measurements to get steady-state transport data for halide perovskites? J. Appl. Phys. 124, 103103 (2018).
- Wehrenfennig, C., Eperon, G. E., Johnston, M. B., Snaith, H. J. & Herz, L. M. High charge carrier mobilities and lifetimes in organolead trihalide perovskites. Adv. Mater. 26, 1584-1589 (2014)
- Sum, T. C. et al. Spectral features and charge dynamics of lead halide perovskites: origins and interpretations. Acc. Chem. Res. 49, 294-302 (2016).
- Rosling, M., Bleichner, H., Jonsson, P. & Nordlander, E. The ambipolar diffusion coefficient in silicon: dependence on excess-carrier concentration and temperature. J. Appl. Phys. 76, 2855-2859 (1994).
- 30. Sinton, R. A. & Cuevas, A. Contactless determination of current-voltage characteristics and minority carrier lifetimes in semiconductors from quasi steady state photoconductance data. Appl. Phys. Lett. 69, 2510-2512 (1996).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

### **Methods**

### Photo-Hall measurement

The experimental setup is shown in Fig. 1a. All measurements were performed at room temperature. Photoexcitation was achieved using a solid-state red laser ( $\lambda$  = 638 nm, maximum power 190 mW) for the perovskite and silicon, or a blue laser ( $\lambda = 450$  nm, maximum power 500 mW) for kesterite (Supplementary Information sections E, G). The sample is centred between the PDL magnets and the laser beam is directed through a motorized neutral density filter. A cylindrical lens is used to expand the beam, while a wedge lens deflects the beam onto the sample area. A beam splitter is used to split the beam towards a 'monitor' silicon photodetector (PD) to measure the monitor photocurrent ( $I_{PD-mon}$ ) at every light intensity. The  $I_{PD-mon}$  is used to determine the incident photon flux  $(\Phi)$  and the absorbed photon density  $(G_v)$  on the sample, which is given as  $G_v = (1 - R)\Phi[1 - \exp(-\alpha d)]/d$ , where R is the reflectivity,  $\alpha$  is the absorption coefficient and d is the thickness (Supplementary Information section C). The optical properties of the films studied in this work are presented in Supplementary Table 2.

The details of the PDL Hall system are described in Supplementary Information section A. The electronic instrumentation consists of a custom-built PDL motor control box, Keithley 2450 source meter unit (SMU) to apply the voltage or current source to the sample, Keithley 2001 digital multimeter (DMM) for voltage measurement, Keithley 7065 Hall switch matrix card with high impedance buffer amplifiers for routeing the signals between the samples, SMU and DMM. For samples with relatively high mobility (perovskite and silicon) we use the d.c. current excitation mode, and for low mobility samples (for example, kesterite) we use a.c. current excitation mode with SRS830 lock-in amplifier to achieve better noise rejection. The PD current is measured using a Keithley 617 electrometer.

At every light intensity, the sheet resistance  $(R_s)$  is obtained by measuring two states and eight states of longitudinal magnetoresistance  $(R_{xx})$  for six-terminal Hall bar and four-terminal van der Pauw samples, respectively. For the Hall bar sample, the sheet resistance is given as  $R_s = R_{xx}W/L$ , where W is the width and L is the length of the Hall bar active area. The conductivity of the sample is given as:  $\sigma = 1/R_s d$ . Next, the transverse magnetoresistance  $(R_{rv})$  is measured. The PDL master magnet is rotated by a stepper motor system, typically with a speed of 1-2 r.p.m., to generate an a.c. field. A typical magnetic field amplitude on the sample is around 0.70 T for a PDL magnet gap of around 10 mm. A Hall sensor is placed under the master magnet to monitor the oscillating field. The field and  $R_{xy}$  are recorded as a function of time for 15-30 min each sweep. This measurement is repeated at several light intensities ranging from dark to the brightest level, while recording the 'monitor' PD current to determine  $G_{v}$ . After all of the measurements are completed, the sample is replaced with a 'reference' photodetector (PD). We then determine the photocurrent ratio,  $k_{PD} = I_{PD-ref}/I_{PD-mon}$ , between the reference PD and the monitor PD at every given light intensity (see Supplementary Information section C for further detail).

Hall signal analysis is performed using a custom program developed in MATLAB<sup>22</sup>. Fourier spectral analysis is used to determine the existence of the magnetoresistance signal ( $R_{xy}$ ) at the same frequency as the magnetic field. We then proceed with phase sensitive lock-in detection, implemented by software, to extract the in-phase component of the Hall signal ( $R_{\rm H}$ ) while rejecting the out-of-phase component that arises from various sources, such as due to Faraday induction of an electromotive force. We use a typical lock-in time constant of 120–300 s. The Hall coefficient is calculated using  $H = R_{\rm H} d/B_0$ , where  $B_0$  is the magnetic field amplitude. Therefore, at every light intensity, we obtain a set of  $\sigma$  and H values and proceed with the photo-Hall analysis. We also provide additional CRPH studies for MAPbI<sub>3</sub> perovskite and a kesterite sample in Supplementary Information sections E, G. For the alternative lifetime measurement in silicon using quasi-steady-state photoconductance technique we used <sup>30</sup> Sinton Instruments WCT-120.

The measurement was performed on the 6" silicon wafer from which the Hall sample originated.

We also make a general remark about the CRPH analysis. First, in the very low light intensity (or low injection) regime—for example,  $G_{\gamma} < 5 \times 10^{17} \, \mathrm{cm}^{-3} \, \mathrm{s}$  for the silicon sample in Fig. 3c—the CRPH analysis results become inaccurate, the  $\Delta n$  values are scattered higher than expected from the monomolecular recombination trend. From our experience, in analysing many samples so far (for example, perovskite, Si, kesterite), we expect that this could be due to the accuracy limitation of the  $\sigma$  and H measurements. For very low light intensity, typically we cannot resolve  $\Delta n$  values smaller than 1% of  $p_0$ , as shown in Fig. 3c. The best analysis results come from the higher intensity regime, in which there is large  $\Delta n$  ( $\Delta n \gg p_0/100$ ) such that substantial changes in  $\sigma$  and H are noted, as in the perovskite example. Second, if the mobility values of the systems are very low, such as in the case of kesterite samples ( $\mu_P \approx 1 \, \mathrm{cm}^2 \, \mathrm{V}^{-1} \, \mathrm{s}^{-1}$ ) and  $\mu_N \approx 10 \, \mathrm{cm}^2 \, \mathrm{V}^{-1} \, \mathrm{s}^{-1}$ ), the Hall coefficient measurement becomes noisier and this also affects the accuracy of the analysis.

### Perovskite solar cell

The perovskite films for photo-Hall study are based on the (FAPbI<sub>3</sub>)<sub>1-x</sub> (MAPbBr<sub>3</sub>)<sub>x</sub> mixed-perovskite system, and use a halide perovskite composition analogous to that used in a previous device with a PCE of 17.9% at x = 0.15 as reported in ref. <sup>24</sup>. The PCE was subsequently improved by modifying the film deposition method and adjusting the value of x. We demonstrated a high average PCE of 20.8% at x = 0.12 for a (FAPbI<sub>3</sub>)<sub>0.88</sub> (MAPbBr<sub>3</sub>)<sub>0.12</sub> device (FTO/bl-TiO<sub>2</sub>/mp-TiO<sub>2</sub>/perovskite/PTAA/Au) under 1 sun conditions (AM 1.5G spectrum, 100 mW cm<sup>-2</sup>). Extended Data Fig. 1a shows photocurrent density-voltage (J-V) curves for the (FAPb  $I_3$ )<sub>0.88</sub>(MAPbBr<sub>3</sub>)<sub>0.12</sub> device measured by reverse and forward scans with 10-mV voltage steps and 40-ms delay times under AM 1.5G illumination. The device exhibits a short-circuit current density  $(J_{SC})$  of 23.3 mA cm<sup>-2</sup>, open circuit voltage ( $V_{\rm OC}$ ) of 1.13 V, and fill factor (FF) of 80.0% by reverse scan. A slightly decreased FF (to 77.8%) by forward scan results in an average PCE of 20.8%. An external quantum efficiency (EQE) spectrum for the device is shown in Extended Data Fig. 1b, demonstrating a very broad plateau of over 80% between 400 nm and 750 nm. The histogram of PCE values for 80 cells is shown in f Fig. 1c.

The J-V curves were measured using a solar simulator (Newport, Oriel Class A, 91195A) with simulated AM1.5G illumination at 100 mW cm<sup>-2</sup> and a calibrated Si-reference cell certificated by the National Renewable Energy Laboratory. The system uses a Keithley 2420 source meter for I-V measurement. The measurement was performed at 25 °C under ambient conditions. The devices were pre-illuminated for 30 s under 1 sun and the measurement was performed in the reverse (from 1.5 V to -0.2 V) and the forward (from -0.2 V to 1.5 V) scanning directions. The current density-voltage (J-V) curves for the perovskite devices were measured by masking the active area (0.16 cm<sup>2</sup> measured using an optical microscope) with a metal mask of 0.094 cm<sup>2</sup> in area. The EQE was measured by a power source (Newport 300W Xenon lamp, 66920) with a monochromator (Newport Cornerstone 260) and a multimeter (Keithley 2001).

### Fabrication of perovskite solar cells

A 70-nm-thick blocking layer of TiO $_2$  (bl-TiO $_2$ ) was deposited onto an F-doped SnO $_2$  (FTO, Pilkington, TEC8) substrate by spray pyrolysis using a 10 vol% titanium diisopropoxidebis (acetylacetonate) solution in ethanol at 450 °C. A TiO $_2$  slurry was prepared by diluting TiO $_2$  pastes (Share Chem, SC-HTO40) in mixed solvent (2-methoxyethan ol:terpineol = 3.5:1 w/w). The 100-nm-thick mesoporous-TiO $_2$  (mp-TiO $_2$ ) was fabricated by spin coating the TiO $_2$  slurry onto the bl-TiO $_2$  layer and subsequently calcining at 500 °C for 1 h in air to remove the organic components. Bis(trifluoromethane) sulfonimide lithium salt was treated onto the mp-TiO $_2$  layer. Then, the (FAPbI $_3$ ) $_{0.88}$ (MAPbBr $_3$ ) $_{0.12}$  film was formed using the method described in the section 'Fabrication of perovskite Hall samples'. A polytriarylamine (PTAA) (EMindex,

 $M_{\rm n}$  = 17,500 g mol<sup>-1</sup>)/toluene (10 mg/1 ml) solution with an additive of 7.5  $\mu$ l Li-bis (trifluoromethanesulfonyl)imide (Li-TFSI)/acetonitrile (170 mg/1 ml) and 4  $\mu$ l 4-*tert*-butylpyridine (TBP) was spin-coated on the perovskite layer/mp-TiO<sub>2</sub>/bl-TiO<sub>2</sub>/FTO substrate at 3,000 r.p.m. for 30 s.

### Fabrication of perovskite Hall samples

All precursor materials were prepared following previous report<sup>24</sup>. To form the perovskite thin film based on the (FAPbI<sub>3</sub>)<sub>0.88</sub>(MAPbBr<sub>3</sub>)<sub>0.12</sub> absorber, the 1.05 M solution dissolving NH<sub>2</sub>CH = NH<sub>2</sub>I(FAI) and CH<sub>3</sub>NH<sub>3</sub>Br(MABr) with PbI<sub>2</sub> and PbBr<sub>2</sub> in N,N-dimethyl formamide (DMF) and dimethyl sulfoxide (DMSO) (6:1 v/v) was prepared by stirring at 60 °C for 1h. Then the solution was coated onto a fused silica substrate heated to 60 °C by two consecutive spin-coating steps, at 1,000 and 5,000 r.p.m., for 5 s and 10 s, respectively. During the second spincoating step, 1 ml ethyl ether was poured onto the substrate after 5 s. Then, the substrate was heat-treated at 150 °C for 10 min. The compact  $(FAPbI_3)_{0.88}(MAPbBr_3)_{0.12}$  film with a thickness of 550 nm was obtained. Then, we selectively scraped the film off the substrate to pattern with the desired Hall bar configuration for photo-Hall measurement. The Hall bar is a six-terminal device as shown in Fig. 2b, inset. We deposited an Au metal contact pattern (100-nm thick) and installed a header pin to mount the sample to the PDL Hall tool.

### **Data availability**

The datasets generated and analysed during the current study are available from the corresponding author on reasonable request.

Acknowledgements S.R.P. and B.S. acknowledge financial support from the Technology Development Program to Solve Climate Changes of the National Research Foundation (NRF) funded by the Ministry of Science, ICT & Future Planning (no. 2016M1A2A2936757), and from the Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Trade, Industry & Energy (MOTIE) of the Republic of Korea (no. 20173010012980). J.H.N. acknowledges financial support from NRF grants funded by the Korea government (MSIP) (2017R1A2B2009676, 2017R1A4A1015022). D.B.M. and O.G. thank the National Science Foundation for support under grant no. DMR-1709294. We thank S. Guha for managing the IBM photovoltaics program; H. Hamann for support; M. Pereira and K. F. Tai for PDL Hall system development; B. Hekmatshoartabari for the silicon sample; and J. Kim for Supplementary Table 4.

Author contributions O.G. and B.S. conceived the project. O.G. led the project, built the experimental setup, programmed the analysis software, derived equation (1) and other formulas, and performed measurement and analyses. S.R.P. prepared samples, and performed optical and Hall measurements and analysis. O.G., S.R.P., B.S. and D.M.B. developed data analysis, interpretation and participated in manuscript writing. Y.V. helped with the development of the PDL system and derivation of the formulae. Y.S.L. and D.M.B. helped with the optical study. N.J.J. and J.H.N. prepared the pervoxkite samples and solar cells. D.B.M., T.T. and X.S. developed the champion CZTSSe process. D.B.M. managed the IBM photovoltaics program and participated in manuscript writing.

Competing interests The PDL Hall system was developed at IBM Research and documented in the following patent families: (1) O. Gunawan & T. Gokmen, US 9,041,389 (ref. <sup>21</sup>); (2) O. Gunawan & M. Pereira, US 9,772,385 (ref. <sup>22</sup>), US 9,678,040, US 15/581183 and related patent applications (WO 201616277241, UK 1717263.6, Japan 2017-552496, Germany 112016000875.9); (3) O. Gunawan, US 15/281,968; (4) O. Gunawan & W. Zhou, US 16/382,937. Patent families (1) and (2) cover the basic a.c. field/PDL Hall system, and (3) and (4) cover the related photo-Hall setup and method.

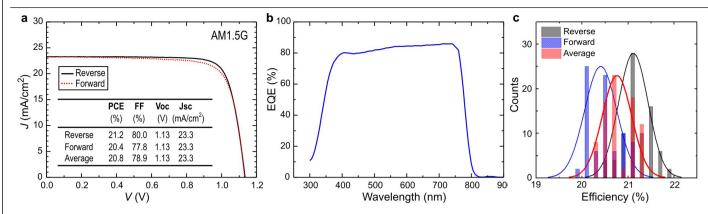
#### Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1632-2

Correspondence and requests for materials should be addressed to O.G. or B.S.

Peer review information *Nature* thanks Henry Snaith and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints.



**Extended Data Fig. 1** | **Performance of the (FAPbl**<sub>3</sub>) $_{0.88}$  (MAPbBr<sub>3</sub>) $_{0.12}$  solar cell **device.** a, Current density-voltage (J-V) curves measured by reverse (black) and forward (red) scans. The photovoltaic performance values are summarized

in the table. **b**, The external quantum efficiency spectrum. **c**, Histogram of the power conversion efficiencies obtained from J-V curves measured by reverse scan (grey) and forward scan (blue), and the average for 80 cells (red).

# High-temperature superconductivity in monolayer Bi<sub>2</sub>Sr<sub>2</sub>CaCu<sub>2</sub>O<sub>8+δ</sub>

https://doi.org/10.1038/s41586-019-1718-x

Received: 2 April 2019

Accepted: 23 August 2019

Published online: 30 October 2019

Yijun Yu<sup>1,2,3,7</sup>, Liguo Ma<sup>1,2,3,7\*</sup>, Peng Cai<sup>1,2,3,7</sup>, Ruidan Zhong<sup>4</sup>, Cun Ye<sup>1,2,3</sup>, Jian Shen<sup>1,2,3</sup>, G. D. Gu<sup>4</sup>, Xian Hui Chen<sup>3,5,6</sup>\* & Yuanbo Zhang<sup>1,2,3</sup>\*

Although copper oxide high-temperature superconductors constitute a complex and diverse material family, they all share a layered lattice structure. This curious fact prompts the question of whether high-temperature superconductivity can exist in an isolated monolayer of copper oxide, and if so, whether the two-dimensional superconductivity and various related phenomena differ from those of their three-dimensional counterparts. The answers may provide insights into the role of dimensionality in high-temperature superconductivity. Here we develop a fabrication process that obtains intrinsic monolayer crystals of the hightemperature superconductor Bi<sub>2</sub>Sr<sub>2</sub>CaCu<sub>2</sub>O<sub>8+6</sub> (Bi-2212; here, a monolayer refers to a half unit cell that contains two CuO<sub>2</sub> planes). The highest superconducting transition temperature of the monolayer is as high as that of optimally doped bulk. The lack of dimensionality effect on the transition temperature defies expectations from the Mermin-Wagner theorem, in contrast to the much-reduced transition temperature in conventional two-dimensional superconductors such as NbSe<sub>2</sub>. The properties of monolayer Bi-2212 become extremely tunable; our survey of superconductivity, the pseudogap, charge order and the Mott state at various doping concentrations reveals that the phases are indistinguishable from those in the bulk. Monolayer Bi-2212 therefore displays all the fundamental physics of high-temperature superconductivity. Our results establish monolayer copper oxides as a platform for studying high-temperature superconductivity and other strongly correlated phenomena in two dimensions.

In systems with reduced dimensions, long-range order (superconductivity in particular) is strongly suppressed<sup>1,2</sup>, as in the case of conventional Bardeen-Cooper-Schrieffer-type superconductors<sup>3,4</sup>, and yet all high-temperature copper oxide superconductors have a layered structure with varying degrees of anisotropy. This apparent dichotomy may be the key to high-temperature superconductivity (HTS)<sup>5-9</sup>, and it raises the question of whether HTS and various correlated phenomena associated with it are different in two dimensions. This question is important for two reasons. First, most HTS theories are based on purely two-dimensional (2D) models<sup>10-12</sup>, whereas experiments show that supercurrent phase coherence<sup>13</sup>, charge ordering<sup>14,15</sup> and charge dynamics<sup>16</sup> all have a 3D nature<sup>17</sup>. Second, much of what we know about HTS came from experimental tools such as scanning tunnelling microscopy/spectroscopy (STM/STS) and angle-resolved photoemission spectroscopy (ARPES) that probe the surface of the materials 18-36; HTS as a bulk property was inferred from the surface measurements. The bulk-surface correspondence becomes ideal if the HTS is truly 2D. To resolve these issues experimentally, an isolated monolayer hightemperature superconductor is needed. Such an atomically thin crystal would represent an ideal correlated 2D system for exploring quantum phenomena in reduced dimensions.

Monolayer HTS has previously been studied mostly in epitaxial oxide heterostructures  $^{37-39}$ , where the active layers are buried between interfaces. Such systems are not accessible to spectroscopic tools such as STM/STS and ARPES. In recent years, an alternative, top-down approach has emerged: it has become possible to mechanically exfoliate monolayer atomic crystals (termed '2D materials') from the layered bulk  $^{40,41}$ . High-quality 2D materials ranging from insulators to metals and superconductors  $^{42}$  have been produced this way.

Experimentally extracting monolayers from bulk high-temperature superconductors, however, turned out to be extremely challenging. Although many of the bulk high-temperature superconductors are considered stable under ambient conditions, they are highly prone to chemical degradation when thinned to monolayers. Indeed, monolayer Bi-2212 has been found to be insulating  $^{41,43}$  or superconducting with a much reduced transition temperature ( $T_{\rm c})^{44}$ . The suppression is seemingly consistent with increased fluctuations expected in 2D superconductors. But given that the material is extremely sensitive

<sup>1</sup>State Key Laboratory of Surface Physics and Department of Physics, Fudan University, Shanghai, China. <sup>2</sup>Institute for Nanoelectronic Devices and Quantum Computing, Fudan University, Shanghai, China. <sup>3</sup>Collaborative Innovation Center of Advanced Microstructures, Nanjing, China. <sup>4</sup>Condensed Matter Physics and Materials Science Department, Brookhaven National Laboratory, Upton, NY, USA. <sup>5</sup>Hefei National Laboratory for Physical Science at Microscale and Department of Physics, University of Science and Technology of China, Hefei, China. <sup>6</sup>Key Laboratory of Strongly Coupled Quantum Matter Physics, University of Science and Technology of China, Hefei, China. <sup>7</sup>These authors contributed equally: Yijun Yu, Liguo Ma, Peng Cai. <sup>\*</sup>e-mail: malq.phys@gmail.com; chenxh@ustc.edu.cn; zhyb@fudan.edu.cn

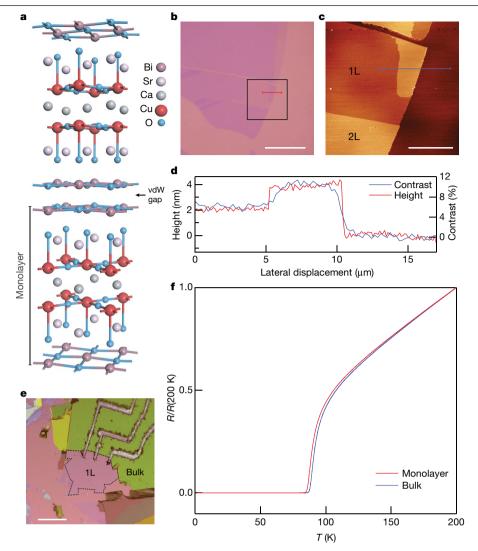


Fig. 1| Fabrication and characterization of atomically thin Bi-2212 transport devices. a, Atomic structure of Bi-2212. 'Monolayer' refers to a half unit cell in the out-of-plane direction that contains two CuO<sub>2</sub> planes. The monolayers are separated by van der Waals gaps in bulk Bi-2212. b, Optical image of a typical Bi-2212 thin flake exfoliated on Si wafer covered with 285-nm-thick SiO<sub>2</sub>. Scale bar. 30 μm. c, Atomic force microscopy (AFM) image of the same flake shown in b (region marked by the black square). L, layer. Scale bar, 10 μm. d, Cross-sectional profile of optical contrast along the red line in **b**, in comparison with the cross-

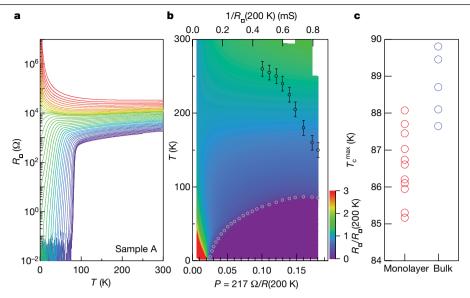
sectional profile of AFM topography at the same location (blue line in c). The quantized steps in contrast and height profiles correspond to monolayer terraces of Bi-2212.e, Optical image of a monolayer Bi-2212 device. The bulk flake in contact with the monolayer is cut into separate pieces, which serve as electrical leads for transport measurements. Scale bar, 100 um. f. Typical temperature-dependent resistance of a monolayer Bi-2212 sample (red) in comparison with that of an optimally doped bulk crystal (blue). Resistances are normalized by their values at  $T = 200 \,\mathrm{K}$ .

to environment and to doping variations, all extrinsic factors must be eliminated before ascribing the reduction of  $T_c$  in monolayers to the effect of dimensionality. The outstanding challenge has been to fabricate high-quality monolayer crystals and probe their intrinsic electronic structure.

Here we overcome these challenges by developing sample fabrication processes that preserve the intrinsic properties of monolayer Bi-2212. We first pinpoint two main causes of sample degradation—reaction with water vapour and rapid loss of oxygen dopant. We find that the degradation slows down in a cold, inert environment, in which pristine monolayer Bi-2212 can be obtained. Unlike the bulk crystal, the monolayer Bi-2212 is extremely tunable: we can continuously vary its doping level in situ and map out major phases from the over-doped regime to the Mott insulating regime, in a single monolayer device. We find that the highest  $T_c$  of the monolayer is as high as that of optimally doped bulk. Moreover, STM/STS study reveals that the monolayer develops the same rich set of phases-HTS, pseudogap, charge order and Mott insulating phase, in particular—that were observed on the bulk surface. Detailed characterization of the phases reveals that they are indistinguishable from those in the bulk. A monolayer, therefore, contains all the essential physics of Bi-2212: that is, HTS in Bi-2212 is essentially a 2D phenomenon.

### Fabricating pristine monolayer Bi-2212

We start with bulk Bi-2212 with a slightly modified stoichiometry,  $Bi_{1.9}Sr_{2.1}CaCu_2O_{8+\delta}$ , which has a highest  $T_c$  of 88 K at optimal doping. In a monolayer Bi-2212, two CuO<sub>2</sub> planes-separated by a Calayer-are sandwiched between SrO and BiO planes to form a charge-neutral, septuplelayered slab as shown in Fig. 1a. The parent compound of Bi-2212 is an antiferromagnetic Mott insulator<sup>45</sup>. Doping holes into the CuO<sub>2</sub> planes generates a pseudogap phase that is characterized by strong depletion of density of states (DOS) near the Fermi level<sup>18–20</sup>. As the doping level p (holes per CuO<sub>2</sub> plaquette) increases, the pseudogap phase evolves



**Fig. 2** | **Tunable high-temperature superconductivity in monolayer Bi-2212. a**, Temperature-dependent resistivity  $R_{\square}(p,T)$  of a monolayer Bi-2212 (sample A) that is initially over-doped. Data were acquired between annealing cycles that progressively lower the doping level of the sample (from purple to red). **b**, Conductivity plotted as a function of temperature and doping level. Doping level p is determined from  $p = 217 \ \Omega/R_{\square}(T = 200 \ \text{K})$ . Black circles denote the onset of the pseudogap state at  $T^*$ . Here the vertical error bars represent uncertainties in locating  $T^*$  at which the temperature-dependent resistance

deviates from linear behaviour. White circles mark the superconducting transition temperature  $T_c$ . The phase diagram spans the optimal doping at which  $T_c$  reaches its maximum value  $T_c^{\text{max}}$ .  $\mathbf{c}$ ,  $T_c^{\text{max}}$  obtained from different monolayer Bi-2212 samples (an example is shown in  $\mathbf{b}$ ), in comparison with  $T_c$  in optimally doped bulk crystals. The highest  $T_c^{\text{max}}$  represents the maximum  $T_c$  of the most intrinsic monolayer in our experiment, and its value lies within the uncertainty range of the  $T_c$  in optimally doped bulk.

into a superconducting phase with highest  $T_c$  reaching 91 K at an optimal doping level of p=0.16 (ref.  $^{46}$ ). Oxygen doping is therefore a key variable that determines the electronic structure in Bi-2212. Because the van der Waals interaction between the layers is weak, atomically thin Bi-2212 flakes can be obtained through mechanical exfoliation on an oxygen-plasma-treated  $SiO_2$  surface  $^{47}$ . Figure 1b and e displays optical images of few-layer Bi-2212 in which the monolayer region is as large as several hundreds of micrometres in diameter (the number of layers is identified from the optical contrast, which correlates well with the thickness of the crystals determined from atomic force microscopy; Fig. 1d).

The exfoliated monolayer Bi-2212 is extremely sensitive to its environment. We find that the monolayers are insulating if the specimen is prepared under ambient conditions, consistent with previous reports 41,43. A systematic investigation (see Extended Data Table 1 and Extended Data Fig. 1) reveals that exposing the monolayers to air, albeit briefly, renders them insulating. Guided by the investigation, we succeeded in obtaining high-quality, intrinsic monolayer Bi-2212 by fabricating samples on a cold stage kept at -40 °C inside an Ar-filled glove box with water and oxygen content below 0.1 ppm. Finally, we make electrical contacts to the monolayer flakes by cold-welding indium/gold microelectrodes (see Methods and Extended Data Table 1) on top. The flakes are then cut into an appropriate geometry with a sharp tip (Fig. 1e), and quickly transferred into an evacuated sample chamber for subsequent transport measurements. We have also obtained monolayer Bi-2212 of similar quality at low temperatures under ultra-high vacuum (UHV) for separate STM/STS study; details of the sample fabrication procedure are provided in the Methods.

Figure 1f shows the normalized resistance of a monolayer in comparison with that of optimally doped bulk Bi-2212. The monolayer retains HTS, and the sharp superconductivity transition signifies the high quality of the sample. More surprisingly, the  $T_{\rm c}$  of the monolayer is almost as high as the optimal  $T_{\rm c}$  in the bulk, indicating that HTS in 2D monolayer Bi-2212 does not differ appreciably from that in 3D bulk. This is corroborated by an accurate quantitative comparison of monolayer and bulk  $T_{\rm c}$ , which we discuss below.

### **Tunable high-temperature superconductivity**

The reduction in dimensionality produces a key advantage: the HTS in monolayer Bi-2212 becomes extremely tunable. The tunability stems from the fact that both sides of the monolayer are exposed, making it easy for interstitial oxygen to escape from or enter the crystal. Specifically, we find that mild vacuum annealing at temperatures between 300 K and 380 K drives oxygen out of the monolayer. Meanwhile, annealing at about 200 K in ozone (partial pressure approximately 0.5 mbar) increases the oxygen concentration (Extended Data Fig. 2). These findings enable us to continuously vary the doping level and track the evolution of various phases, including superconductivity, from an over-doped to deeply under-doped regime (and vice versa) in a single monolayer sample. Figure 2a displays a set of measurements of temperature-dependent resistivity,  $R_{\Box}(T)$ , of a monolayer Bi-2212 (sample A), acquired between annealing treatments at 300–380 K in vacuum (base pressure <10<sup>-4</sup> mbar). The annealing treatments progressively lower the hole doping level in the monolayer and induce a transition from superconducting to insulating behaviour. Meanwhile, the room-temperature resistivity increases by one order of magnitude from about  $1\,k\Omega$  to about  $30\,k\Omega$ . Details of the transition become more apparent when the resistivity of the same sample (normalized to its value at T = 200 K) is plotted as a function of temperature and hole doping level p, as shown in Fig. 2b. (Here the hole doping level is determined from  $p = \text{const.}/R_{\Box}(T = 200 \text{ K})$  the value of the constant (const.) is chosen so that p = 0.16 at optimal doping <sup>48,49</sup>, and the precise value of p does not affect our conclusions.) As p decreases,  $T_c$  (defined as the temperature at which  $d^2R_{\Box}/dT^2 = 0$ ; see Extended Data Fig. 3) rises at first, then falls continuously, giving rise to a superconducting dome that ends at  $p \approx 0.022$ . An insulating phase appears next to the superconducting dome. In addition, we observe at  $T^* > T_c$  the onset of the pseudogap phase that is marked by deviation from a linear  $R_{\Box}(T)$  in the normal state of a high-temperature superconductor under various doping levels (open black circles in Fig. 2b; see Extended Data Fig. 3 for detailed analysis). Figure 2b, therefore, maps out a phase

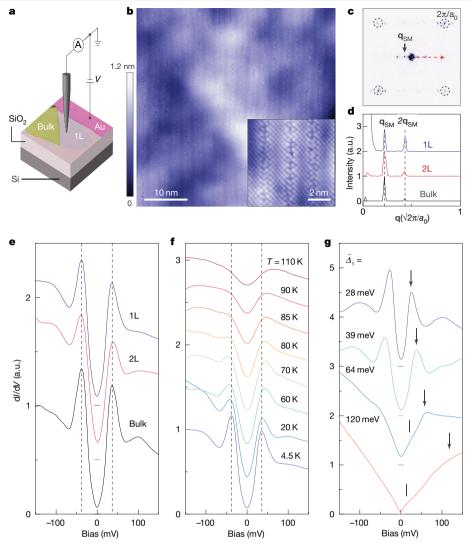
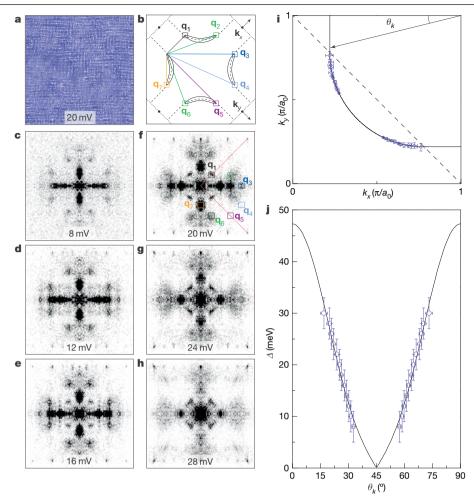


Fig. 3 | Tunnelling spectroscopy of monolayer Bi-2212. a, Schematic illustration of the STM measurement set-up. Monolayer Bi-2212 (and the bulk crystal nearby) is electrically connected to a pre-patterned Au electrode, which provides a returning path for the tunnelling current. b, High-resolution STM topograph of a monolayer Bi-2212. The image was taken at a junction resistance of  $2 G\Omega$  and a sample bias voltage of -300 mV. Inset: magnified view of top Bi atoms and supermodulation ridges. c, Fourier transform of the STM topograph in **b**. Peaks are clearly visible at multiples of supermodulation wavevector  $\mathbf{q}_{\text{SM}}$ . Bragg peaks of the atomic lattice are marked by broken circles. **d**, Line cut of the Fourier transform in **c** along the [1,1] direction.  $Monolayer\,curve\,(1L,blue)\,is\,compared\,with\,bilayer\,(2L,red)\,and\,bulk\,(black)$ data; peaks at  $\mathbf{q}_{SM}$  and  $2\mathbf{q}_{SM}$  align within an uncertainty of 1.5% of  $2\pi/a_0$ .

 $\pmb{e}, Spatially \, averaged \, differential \, conductance \, spectra \, acquired \, on \, monolayer,$ bilayer and bulk Bi-2212 at near optimal doping. Broken lines mark the position of coherence peaks. The horizontal bars mark the zero of each curve. f, Spatially averaged spectral temperature dependence on a nearly optimal doped monolayer Bi-2212 showing a smooth transition from the superconducting to pseudogap state at  $T \approx 85$  K. g, Evolution of spatially averaged tunnelling spectra of monolayer Bi-2212 with diminishing doping level p. Here the doping level is characterized by  $\bar{\Delta}_1(p)$ . The energies  $\Delta_1$  (black arrows) are extracted at the pseudogap edge, and the energies  $\Delta_0$  (vertical bars) are identified as the 'kink' energy<sup>23</sup>. Curves are offset for clarity, and horizontal bars mark the zero of each curve.

diagram of the monolayer that is strikingly similar to that of bulk copper oxides50.

Close examination of the phase diagram in Fig. 2b provides further insights into the 2D HTS in monolayer Bi-2212. We focus on the high  $T_c$ that characterizes the superconducting transition in the monolayer. Specifically, we use the phase diagram to accurately determine how much, if at all,  $T_c$  is suppressed in the monolayer compared with in the bulk. Because  $T_c$  strongly depends on hole doping level, a comparison is valid only when it is made at the same doping level. The maximum  $T_c$ at optimal doping,  $T_c^{\text{max}}$ , therefore serves as a natural metric for such comparison, given that varying the sample thickness does not alter the optimal doping level itself. Figure 2c summarizes the measured  $T_c^{\text{max}}$  of monolayer Bi-2212 in comparison with the  $T_c$  of optimally doped bulk crystals. (Here  $T_{\rm c}^{\rm max}$  of monolayers was extracted from phase diagrams, exemplified in Fig. 2b, and we ensured that the superconducting domes of all monolayer samples spanned the optimal doping so that  $T_c^{\text{max}}$ could be reliably determined;  $T_c^{\text{max}}$  determined by different methods is shown in Extended Data Fig. 3.) Both datasets exhibit appreciable spread that most likely reflects variations in the impurity level in different specimens. More importantly, the highest  $T_{\rm c}^{\rm max}$  of 88.1 K that represents the most intrinsic monolayer is within the uncertainty of optimal bulk  $T_c$ . The difference of about 2% between the average of  $T_c^{\text{max}}$ in the monolayer and the average of optimal  $T_c$  in bulk may be explained by inevitable slight sample degradation from our fabrication process. Our observations therefore reveal a robust 2D HTS in monolayer Bi-2212 with optimal transition temperature as high as that in 3D bulk.



**Fig. 4** | **Quasi-particle interference and superconducting gap in monolayer Bi-2212. a**, Representative conductance ratio map  $Z(\mathbf{r}, E)$  obtained at E = 20 meV on the same area as in Fig. 3b. **b**, Illustration of the octet model for Bogoliubov quasiparticle interference in Bi-2212 at a given energy. The octet ends of four banana-shaped constant-energy contours have maximum density of states. Quasi-particle scattering between these eight regions produces seven primary scattering  $\mathbf{q}$ -vectors,  $\mathbf{q}_1$  to  $\mathbf{q}_7$ , labelled by coloured squares.  $\mathbf{c}$ - $\mathbf{h}$ , Fourier transform of the conductance ratio map  $|Z(\mathbf{q}, E)|$ . The Fourier transforms are mirror-symmetrized and normalized to their average value. E is labelled on each panel. In particular,  $\mathbf{f}$  displays the Fourier transform of the conductance ratio map in  $\mathbf{a}$ . Red solid lines indicate the atomic Bragg vectors at  $(2\pi/a_0, 0)$  and  $(0, 2\pi/a_0)$ . Of the total of seven independent scattering vectors (coloured

squares) prescribed by the octet model illustrated in  $\mathbf{b}$ , five are observed as peaks in the Fourier transform;  $\mathbf{q}_4$  and  $\mathbf{q}_5$  are too weak to be detected.  $\mathbf{i}$ , Loci of the ends of banana-shaped constant-energy contours extracted from dispersion of the  $\mathbf{q}$ -vectors. Locations of the loci represent the underlying Fermi surface. Solid line is a fit to the data with a circular arcjoined with two straight lines. Broken line marks the antiferromagnetic zone boundary.  $\mathbf{j}$ , Superconducting gap  $\Delta_{SC}$  as a function of Fermi surface angle  $\theta_k$ .  $\Delta_{SC}$  is extracted from the measured position of scattering vectors  $\mathbf{q}_1$  to  $\mathbf{q}_7$  (excluding  $\mathbf{q}_4$  and  $\mathbf{q}_5$ ) following the procedure described in refs.  $^{24,26}$ . Solid line is a fit to the data with d-wave gap function  $\Delta(\theta_k) = \Delta_{QPI}[A\cos(2\theta_k) + (1-A)\cos(6\theta_k)]$ , where  $\Delta_{QPI} = 47.3$  meV and A = 0.844 are fitting parameters.

### Monolayer topography and tunnelling spectroscopy

STM topography measurement (schematic set-up shown in Fig. 3a) confirms the high quality of monolayer Bi-2212, which retains the original atomic structure found in the bulk crystals. Figure 3b displays the atom-resolved topography of the top BiO plane of a Bi-2212 monolayer. The surfaces are as clean as the bulk surface and are continuous over macroscopic distances (about 100 µm; Extended Data Fig. 6). Nearly commensurate supermodulation ridges along the [110] direction—a distinctive feature in Bi-based bulk copper oxides<sup>18</sup>—are clearly observed. Fourier transform of the topography images reveals that the period of the supermodulation  $\mathbf{q}_{\text{SM}}$  exactly matches that on the bulk surface (Fig. 3c, d); no additional surface reconstructions were detected. Despite the identical atomic structure, monolayer Bi-2212 does exhibit a feature not seen on the bulk surface: large scale corrugations with a root-mean-square (r.m.s.) value of 0.2 nm, in contrast to the flat surface of the bulk crystal. We attribute the corrugations to the underlying substrate: few-layer Bi-2212 may become flexible and partially conform to the rough surface of amorphous  $SiO_2$  (r.m.s. approximately 0.25 nm).

We now turn to the electronic structure of monolayer Bi-2212. We note that a variety of spectroscopy studies revealed a rich set of phases that are characterized by two energy scales, referred to as  $\Delta_0$  and  $\Delta_1$ , in bulk Bi-2212 (refs.  $^{20,21}$ ). Specifically, excitations in the superconducting state occur at energies  $E \lesssim \Delta_0$ , whereas charge-order and other highly correlated broken-symmetry states appear at pseudogap energy scale  $E \approx \Delta_1$ ; the competition or cooperation between these intertwined phases remains one of the central problems of HTS (refs.  $^{5,12}$ ). In the following, we examine these strongly correlated states in monolayer Bi-2212.

Figure 3e displays the differential conductance spectra g(E), which is proportional to the DOS at energy E, of monolayer and bilayer samples cleaved from a nearly optimally doped bulk crystal with  $T_c = 88 \text{ K}$  (referred to as OP88). Here the spectra are spatial averages of the local differential conductance spectra  $g(\mathbf{r}, E = eV) \equiv dI/dV|_{\mathbf{r}, V}$  over a 500 Å

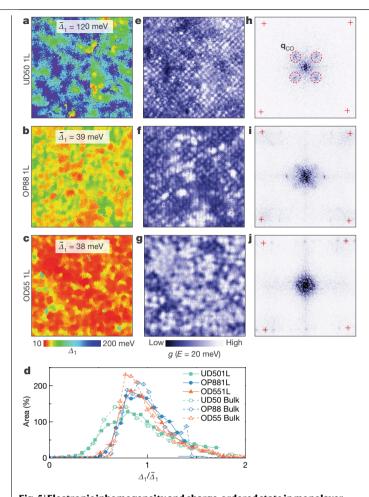


Fig. 5 | Electronic inhomogeneity and charge-ordered state in monolayer **Bi-2212.**  $\mathbf{a}$ - $\mathbf{c}$ , Gap maps  $\Delta_1(\mathbf{r})$  obtained on monolayer Bi-2212. The monolayers are obtained from bulk crystals UD50 (under-doped,  $T_c = 50 \,\mathrm{K}$ ), OP88 (optimally doped,  $T_c = 88 \text{ K}$ ) and OD55 (over-doped,  $T_c = 55 \text{ K}$ ). Field of view,  $400 \text{ Å} \times 400 \text{ Å}$ .  $\bar{\Delta}_1$  denotes the average value of  $\Delta_1$  over the entire field of view.  $\Delta_1(\mathbf{r})$  in **a** was determined from fitting each local tunnelling spectrum using the method described in ref. <sup>36</sup>. Values of  $\Delta_1(\mathbf{r})$  in **b** and **c** were extracted as the energy separation between two coherence peaks in each local tunnelling spectrum. **d**, Histograms of  $\Delta_1(\mathbf{r})$  shown in  $\mathbf{a}-\mathbf{c}$  normalized by their mean value. The normalized gap distributions in monolayers are highly similar to those of bulk source crystals (Extended Data Fig. 9). e-g, Conductance maps  $g(\mathbf{r}, E) = dI/dV(\mathbf{r}, E)$  recorded at E = 20 meV on the same areas shown in  $\mathbf{a} - \mathbf{c}$ .  $\mathbf{h}$ - $\mathbf{j}$ , Fourier transforms of  $g(\mathbf{r}, E = 20 \text{ meV})$  in  $\mathbf{e}$ - $\mathbf{g}$ . Charge-order peaks are clearly resolved at  $\mathbf{q} = (\pm 0.25, 0)2\pi/a_0$  and  $(0, \pm 0.25)2\pi/a_0$  (marked by broken circles) in under-doped monolayer. Red crosses mark lattice wavevectors at  $(\pm 2\pi/a_0, 0)$ and  $(0, \pm 2\pi/a_0)$ .

× 500 Å field of view; I and V are tunnelling current and sample-bias voltage, respectively, and e is the charge of an electron. The V-shaped superconducting energy gap and the large coherence peaks on both sides of the gap are clearly observed in the spectra. The size of the gap, defined as half the separation between two coherence peaks,  $\Delta_0$ , in the monolayer and bilayer is almost identical to that in the bulk (Fig. 3e, black curve) from which the monolayer and bilayer were obtained. Close examination reveals that the monolayer and bilayer spectra also faithfully reproduce the fine details, the dip-hump structure outside of the gap and the electron-hole asymmetric background in particular, that are found in the bulk spectrum<sup>18</sup>. Differential conductance spectra at elevated temperatures show that the pseudogap state, too, persists in monolayer Bi-2212. The pseudogap state manifests as a gap in g(E) well above the  $T_c$  of the bulk source crystal (Fig. 3f). Finally, we note that  $\Delta_1$ coincides with  $\Delta_0$  in the nearly optimally doped monolayer. On lowering the doping level, however, the two energy scales diverge:  $\Delta_1$  moves to higher energies, whereas  $\Delta_0$  becomes smaller (Fig. 3g), consistent with the behaviour in bulk copper oxide superconductors<sup>18,23</sup>. The close match between the monolayer and bulk spectra is the first indication that the superconducting state (and electronic structures associated with it) remains intact in the 2D limit.

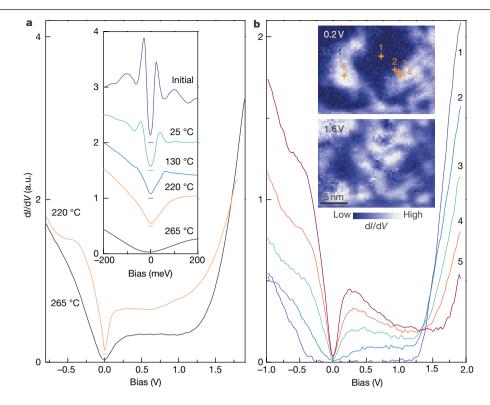
### Quasi-particle interference and superconducting gap

The low-energy excitations inside the superconducting energy gap carry crucial information on the superconducting state. The excitations, also known as Bogoliubov quasiparticles, scatter off impurities and produce interference patterns that can be detected by spatial mapping of the tunnelling conductance in  $g(\mathbf{r}, eV)$  at a given bias V on the bulk Bi-2212 surface<sup>21,24</sup>. Further, the Fourier transform of the interference patterns reveals maxima at a set of energy-dependent wavevectors  $\mathbf{q}_i$  (i=1,...,7)—a result of elastic scattering between the eight high jointdensity-of-state loci of the 'banana-shaped' constant energy contour of Bogoliubov quasiparticles<sup>24</sup> (referred to as the 'octet model'; Fig. 4b). The quasi-particle interference has therefore been a powerful tool for reconstructing the superconducting gap dispersion  $\Delta(\mathbf{k})$  of copper  $oxide \, superconductors^{18,26}.$ 

We used the quasi-particle interference technique to probe  $\Delta(\mathbf{k})$  in monolayer Bi-2212. We focus on the conductance ratio map  $Z(\mathbf{r}, E = eV) \equiv g(\mathbf{r}, +eV)/g(\mathbf{r}, -eV)$ , which eliminates systematic errors related to the tunnelling setpoint associated with directly mapping the conductance  $g(\mathbf{r}, eV)$  (ref. <sup>26</sup>). Figure 4a displays an example of the conductance ratio map of monolayer Bi-2212 obtained at E = 20 meV. The Fourier transform of the conductance ratio map,  $|Z(\mathbf{q}, E = eV)|$  shows clear maxima at q, that are fully consistent with the octet model, except that peaks at  $\mathbf{q}_4$  and  $\mathbf{q}_5$  are too weak to be detected (Fig. 4f). As the tunnelling bias V is varied, we observe that the measured  $\mathbf{q}_i$  disperse with energy E = eV, and the dispersions  $\mathbf{q}_i(E)$  are again consistent with those expected from the octet model (Extended Data Figs. 7 and 8). The dispersions  $\mathbf{q}_i(E)$  allow us to extract the energy-dependent locations of the octet ends of the 'bananas' in k space, and the obtained loci can be interpreted as the normal-state Fermi surface<sup>24</sup>. Our result, shown in Fig. 4i, is consistent with a cylindrical Fermi surface centred at  $(\pi,\pi)$  that is observed in bulk Bi-2212 and various other bulk copper oxide superconductors<sup>26</sup>. Finally, we determine the superconducting gap dispersion  $\Delta(\mathbf{k})$  from  $\mathbf{q}_i(E)$ . Figure 4j displays the measured superconducting gap energy of the monolayer as a function of  $\theta_{\nu}$  along the Fermi surface. The data agree with the d-wave superconducting gap dispersion of bulk Bi-2212 at similar doping level<sup>23</sup>. We therefore conclude that reducing the material's dimensions from three to two does not fundamentally alter the superconducting gap structure.

### Electronic inhomogeneity and charge-ordered state

Next, we focus on the electronic structure of monolayer Bi-2212 beyond the superconducting energy gap  $\Delta_0$ . In particular, the energy scale  $\Delta_1$ is associated with the anti-nodal pseudogap and other correlated states that are intricately linked to superconductivity<sup>21</sup>. In contrast to the relatively homogeneous superconducting gap  $\Delta_0$ , the pseudogap  $\Delta_1$ varies widely at the nanometre length scale on bulk copper oxides<sup>18</sup>. To study the inhomogeneity in monolayer Bi-2212, we extract  $\Delta_1$  from the local differential conductance spectra collected on a dense array of locations on samples at various doping levels, and construct the gap map  $\Delta_1(\mathbf{r})$  as shown in Fig. 5a-c. Similar to previous measurements on bulk Bi-2212, we find that wide, nanometre-scale variations in  $\Delta_1$  diminish as the doping level increases in the monolayer; meanwhile  $\Delta_1$ averaged over the entire field of view,  $\bar{\Delta}_1$ , shifts to lower energies. Close examination of  $\Delta_1$  histograms reveals that  $\bar{\Delta}_1$  in monolayers is in general larger than that in bulk source crystals from which the monolayers are cleaved (Extended Data Fig. 9), and the deviation varies from



**Fig. 6** | **Electronic structure of monolayer Bi-2212 in the Mott insulating regime. a**, Spatially averaged differential conductance spectra of monolayer Bi-2212 obtained between vacuum annealing cycles. The annealing temperature is marked on each curve. The spectrum labelled 'Initial' was recorded before annealing. The as-exfoliated monolayer (obtained from OD55 crystal) was initially over-doped. The annealing cycles progressively lower its

doping level and eventually make the specimen extremely under-doped.  $\boldsymbol{b},$  Representative tunnelling spectra of the extremely under-doped monolayer in  $\boldsymbol{a}.$  Inset: tunnelling conductance maps recorded at tunnelling biases of  $0.2\,V$  (upper panel) and  $1.6\,V$  (lower panel). Crosses mark the positions where the spectra are taken. Spectra are shifted vertically for clarity.

sample to sample. Such deviation is consistent with results from transport measurements; we attribute it to slight loss of oxygen doping (up to 3% in over-doped samples) during sample fabrication. The gap distributions in monolayer and bulk, however, converge if  $\Delta_1$  is normalized to  $\bar{\Delta}_1$  in each gap map (Fig. 5d). This observation suggests that the microscopic mechanism of the  $\Delta_1$  disorder remains the same in the monolayer, even though the monolayer's dielectric environment is, in absence of the interlayer Coulomb interaction, very different from the bulk.

Despite the large spatial inhomogeneity at high energy scale, a periodic chequerboard charge order emerges outside of the superconducting energy gap in various bulk copper oxides<sup>12,28,29</sup>. Recent experiments show mounting evidence that a periodic modulation of Cooper pair $ing-that is, a pair density wave-may coexist with the charge order {}^{12,31,32}.\\$ These charge-ordered states are intimately related to the superconductivity in the CuO<sub>2</sub> plane<sup>12,29</sup>. An important question is then whether these states persist in the 2D limit. Our conductance mapping of an underdoped monolayer answers the question in the affirmative. As shown in Fig. 5e, a chequerboard pattern is resolved on the conductance map  $g(\mathbf{r}, E)$  obtained at E = 20 meV. Fourier transform of the map (Fig. 5h) shows that the chequerboard pattern corresponds to wavevector  $\mathbf{q}_{co}$ around 1/4 of the lattice wavevector  $2\pi/a_0$  along the Cu–Cu bond direction ( $a_0$  is the distance between neighbouring Cu atoms). The CO therefor e has a real-space wavelength of about  $4a_0$ , with a correlation length of about 14  $a_0$  obtained from a Gaussian fit to its peak profile (Extended Data Fig. 10). These results agree well with bulk values 28,29,34. As the doping level increases, the CO diminishes and eventually disappears in the over-doped regime (Fig. 5i, j), consistent with observations in bulk copper oxides<sup>29</sup>. Finally, we present evidence that pair density

waves also exist in monolayer Bi-2212. Here we examine spatial variation in the amplitude of the coherence peak in the tunnelling spectrum, which empirically correlates with Cooper-pair density modulation in bulk Bi-2212, using the procedure described in ref.  $^{32}$ . The coherence peak amplitude map (of the same area as in Fig. 5a, e; Extended Data Fig. 11) exhibits a chequerboard pattern with a period of about  $4a_0$ —a clear signature of a pair density wave order.

### Electronic structure in the Mott insulating regime

Because oxygen in monolayer Bi-2212 escapes easily at elevated temperatures, we are able to access a wide, continuous doping range in a single specimen by gentle annealing in ultra-high vacuum (Fig. 6a and Extended Data Table 2). Here we focus on the extremely underdoped regime, where the pseudogap and charge-ordered states start to emerge from the parent Mott insulator 10,34. Figure 6b displays typical tunnelling spectra obtained on an extremely under-doped monolayer (see inset). The evolution of the spectra is strikingly similar to that in severely under-doped bulk copper oxides<sup>34,51</sup>. A large charge transfer gap of 1.2 eV is observed on Mott insulating patches. (The gap value is 20% larger than that in bulk Bi-2212 (ref. 51); we attribute the discrepancy to the tip-induced band-bending effect that is common in tunnelling spectroscopy studies of insulators<sup>35</sup> and 2D materials<sup>52</sup>.) Outside the Mott insulating patches, a broad in-gap state develops within the charge transfer gap, giving rise to a pseudogap-like spectra around the Fermi level. As in the bulk, the conductance maps at low bias and high bias are anticorrelated (Fig. 6b inset), which implies that the in-gap state comes from spectral weight transfer from the upper Hubbard band of the parent Mott insulator. Our results on monolayers, therefore,

indicate that the dimensionality effect, if it exists at all, does not play an important role in the transition from Mott to pseudogap phase

### **Online content**

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1718-x.

- Mermin, N. D. & Wagner, H. Absence of ferromagnetism or antiferromagnetism in one- or 1. two-dimensional isotropic Heisenberg models. Phys. Rev. Lett. 17, 1133-1136 (1966).
- 2. Kosterlitz, J. M. & Thouless, D. J. Ordering, metastability and phase transitions in two-dimensional systems. J. Phys. C 6, 1181-1203 (1973).
- Saito, Y., Nojima, T. & Iwasa, Y. Highly crystalline 2D superconductors. Nat. Rev. Mater. 2, 3. 16094 (2017).
- Uchihashi, T. Two-dimensional superconductors with atomic-scale thickness, Supercond. 4 Sci. Technol. 30, 013002 (2017).
- Keimer, B., Kivelson, S. A., Norman, M. R., Uchida, S. & Zaanen, J. From quantum matter to 5 high-temperature superconductivity in copper oxides. Nature 518, 179-186 (2015).
- 6. Chakravarty, S., Sudbø, A., Anderson, P. W. & Strong, S. Interlayer tunneling and gap anisotropy in high-temperature superconductors. Science 261, 337-340 (1993)
- 7. Anderson, P. W. Interlayer tunneling mechanism for high-T<sub>c</sub> superconductivity: comparison with c axis infrared experiments. Science 268, 1154-1155 (1995).
- 8. Leggett, A. J. WHERE is the energy saved in cuprate superconductivity? J. Phys. Chem. Solids 59, 1729-1732 (1998).
- Kresin, V. Z. & Morawitz, H. Layer plasmons and high-T<sub>c</sub> superconductivity. Phys. Rev. B 37, 7854-7857 (1988)
- Lee, P. A. & Wen, X.-G. Doping a Mott insulator: physics of high-temperature superconductivity. Rev. Mod. Phys. 78, 17-85 (2006).
- Scalapino, D. J. A common thread: the pairing interaction for unconventional superconductors. Rev. Mod. Phys. 84, 1383-1417 (2012).
- Fradkin, E., Kivelson, S. A. & Tranquada, J. M. Colloquium: Theory of intertwined orders in high temperature superconductors. Rev. Mod. Phys. 87, 457-482 (2015).
- Rajasekaran, S. et al. Probing optically silent superfluid stripes in cuprates. Science 359, 575-579 (2018).
- Gerber, S. et al. Three-dimensional charge density wave order in  $YBa_2Cu_3O_{6.67}$  at high magnetic fields, Science 350, 949-952 (2015).
- Bluschke, M. et al. Stabilization of three-dimensional charge order in YBa<sub>2</sub>Cu<sub>3</sub>O<sub>6+</sub>, via epitaxial growth, Nat. Commun. 9, 2978 (2018).
- Hepting, M. et al. Three-dimensional collective charge excitations in electron-doped
- copper oxide superconductors, Nature 563, 374 (2018) 17. Schneider, T. Dimensional crossover in cuprate superconductors. Z. Phys. B Condens.
- Matter 85, 187-195 (1991). Fischer, Ø., Kugler, M., Maggio-Aprile, I., Berthod, C. & Renner, C. Scanning tunneling spectroscopy of high-temperature superconductors. Rev. Mod. Phys. 79, 353-419 (2007).
- 19. Damascelli, A., Hussain, Z. & Shen, Z.-X. Angle-resolved photoemission studies of the cuprate superconductors. Rev. Mod. Phys. 75, 473-541 (2003).
- Timusk, T. & Statt, B. The pseudogap in high-temperature superconductors: an experimental survey. Rep. Prog. Phys. 62, 61-122 (1999).
- Schmidt, A. R. et al. Electronic structure of the cuprate superconducting and pseudogap
- phases from spectroscopic imaging STM. New J. Phys. 13, 065014 (2011). McElroy, K. et al. Atomic-scale sources and mechanism of nanoscale electronic disorder
- in Bi<sub>2</sub>Sr<sub>2</sub>CaCu<sub>2</sub>O<sub>8+δ</sub>. Science **309**, 1048–1052 (2005).
- Kohsaka, Y. et al. How Cooper pairs vanish approaching the Mott insulator in  $Bi_2Sr_2CaCu_2O_{8+\delta}$ . Nature **454**, 1072–1078 (2008).
- McElroy, K. et al. Relating atomic-scale electronic phenomena to wave-like quasiparticle states in superconducting  $Bi_2Sr_2CaCu_2O_{8+\delta}$ . Nature **422**, 592–596 (2003).
- Hoffman, J. E. et al. Imaging quasiparticle interference in Bi<sub>2</sub>Sr<sub>2</sub>CaCu<sub>2</sub>O<sub>925</sub>, Science 297. 1148-1151 (2002).

- Hanaguri, T. et al. Quasiparticle interference and superconducting gap in Ca2-xNaxCuO2Cl2. Nat. Phys. 3, 865-871 (2007).
- 27 Lang, K. M. et al. Imaging the granular structure of high- $T_c$  superconductivity in underdoped Bi<sub>2</sub>Sr<sub>2</sub>CaCu<sub>2</sub>O<sub>8+δ</sub>. Nature 415, 412 (2002).
- Hanaguri, T. et al. A 'checkerboard' electronic crystal state in lightly hole-doped Ca<sub>2-x</sub>Na<sub>x</sub>CuO<sub>2</sub>Cl<sub>2</sub>. Nature **430**, 1001-1005 (2004).
- daSilva Neto, E. H. et al. Ubiquitous interplay between charge ordering and hightemperature superconductivity in cuprates, Science 343, 393-396 (2014)
- Comin, R. et al. Charge order driven by Fermi-arc instability in  $Bi_2Sr_{2-x}La_xCuO_{6+\delta}$ . Science **343**. 390-392 (2014).
- Hamidian, M. H. et al. Detection of a Cooper-pair density wave in  $Bi_2Sr_2CaCu_2O_{8+x}$ . Nature **532**, 343-347 (2016).
- Ruan, W. et al. Visualization of the periodic modulation of Cooper pairing in a cuprate superconductor, Nat. Phys. 14, 1178 (2018).
- Mesaros, A. et al. Commensurate 4a<sub>0</sub>-period charge density modulations throughout the Bi<sub>2</sub>Sr<sub>2</sub>CaCu<sub>2</sub>O<sub>8+x</sub> pseudogap regime. Proc. Natl Acad. Sci. USA 113, 12661-12666 (2016).
- Cai, P. et al. Visualizing the evolution from the Mott insulator to a charge-ordered insulator in lightly doped cuprates, Nat. Phys. 12, 1047-1051 (2016).
- 35. Kohsaka, Y. et al. Visualization of the emergence of the pseudogap state and the evolution to superconductivity in a lightly hole-doped Mott insulator. Nat. Phys. 8, 534-538 (2012).
- Alldredge, J. W. et al. Evolution of the electronic excitation spectrum with strongly diminishing hole density in superconducting  $Bi_2Sr_2CaCu_2O_{8+\delta}$ . Nat. Phys. 4, 319–326 (2008).
- Gozar, A. et al. High-temperature interface superconductivity between metallic and insulating copper oxides. Nature 455, 782-785 (2008).
- Bollinger, A. T. & Božović, I. Two-dimensional superconductivity in the cuprates revealed by atomic-layer-by-layer molecular beam epitaxy. Supercond. Sci. Technol. 29, 103001 (2016).
- Terashima, T. et al. Superconductivity of one-unit-cell thick YBa<sub>2</sub>Cu<sub>3</sub>O<sub>7</sub> thin film. Phys. Rev. Lett. 67, 1362-1365 (1991).
- Frindt, R. F. Superconductivity in ultrathin NbSe<sub>2</sub> layers. Phys. Rev. Lett. 28, 299-301
- Novoselov, K. S. et al. Two-dimensional atomic crystals. Proc. Natl Acad. Sci. USA 102, 10451-10453 (2005).
- Ajayan, P., Kim, P. & Banerjee, K. Two-dimensional van der Waals materials. Phys. Today 69, 38-44 (2016).
- Sandilands, L. J. et al. Origin of the insulating state in exfoliated high-T<sub>c</sub> two-dimensional atomic crystals, Phys. Rev. B 90, 081402(R) (2014).
- 44. Jiang, D. et al. High-T<sub>c</sub> superconductivity in ultrathin Bi<sub>2</sub>Sr<sub>2</sub>CaCu<sub>2</sub>O<sub>8+x</sub> down to half-unitcell thickness by protection with graphene, Nat. Commun. 5, 5708 (2014).
- Zaanen, J., Sawatzky, G. A. & Allen, J. W. Band gaps and electronic structure of transitionmetal compounds. Phys. Rev. Lett. 55, 418-421 (1985).
- Presland, M. R., Tallon, J. L., Buckley, R. G., Liu, R. S. & Flower, N. E. General trends in 46. oxygen stoichiometry effects on  $T_{\rm c}$  in Bi and Tl superconductors. Physica C 176, 95–105 (1991).
- Huang, Y. et al. Reliable exfoliation of large-area high-quality flakes of graphene and other two-dimensional materials. ACS Nano 9, 10612-10620 (2015)
- Bollinger, A. T. et al. Superconductor-insulator transition in La<sub>2-x</sub>Sr<sub>x</sub>CuO<sub>4</sub> at the pair quantum resistance. Nature 472, 458-460 (2011).
- Leng, X., Garcia-Barriocanal, J., Bose, S., Lee, Y. & Goldman, A. M. Electrostatic control of the evolution from a superconducting phase to an insulating phase in ultrathin YBa<sub>2</sub>Cu<sub>3</sub>O<sub>7-x</sub> films. Phys. Rev. Lett. 107, 027001 (2011).
- Ando, Y., Komiya, S., Segawa, K., Ono, S. & Kurita, Y. Electronic phase diagram of high- $T_{\rm c}$ cuprate superconductors from a mapping of the in-plane resistivity curvature. Phys. Rev. Lett. 93, 267001 (2004).
- Ruan, W. et al. Relationship between the parent charge transfer gap and maximum transition temperature in cuprates. Sci. Bull. (Beijing) 61, 1826-1832 (2016).
- Brar, V. W. et al. Gate-controlled ionization and screening of cobalt adatoms on a graphene surface. Nat. Phys. 7, 43-47 (2011).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

### Methods

### Fabricating monolayer Bi-2212 for transport measurements

Monolayer flakes of Bi-2212 can be obtained through mechanical exfoliation  $^{41}$ . However, the size of thin flakes tends to be small because brittle Bi-2212 crystals break easily during exfoliation. Activating the SiO2 surface with oxygen plasma treatment greatly increases the area and yield of monolayer crystals on the SiO2/Si wafer  $^{47}$ . We attribute the improvement to the enhanced adhesion between Bi-2212 and SiO2—the plasma treatment functionalizes the SiO2 surface with hydroxyl groups that strongly bind to Bi-2212 (ref.  $^{53}$ ).

Our systematic investigation (Extended Data Table 1 and Extended Data Fig. 1) reveals that exposing the monolayers to air, albeit briefly, renders them insulating  $^{41,43,54,55}$ . An inert Ar atmosphere preserves the superconductivity in the monolayers, but the protection is incomplete:  $T_c$  is much suppressed after a prolonged fabrication process at room temperature, and shortening the fabrication time leads to a higher  $T_c$ . These observations point to (1) reaction with water vapour in air and (2) rapid oxygen loss at room temperature as two main causes of degradation in monolayer Bi-2212. (The same degradation pathways are also present in bulk crystals  $^{56-58}$ .) The oxygen loss, however, slows down considerably at moderately low temperatures, so we are able to obtain high-quality, intrinsic monolayers by fabricating samples on a cold stage kept at  $^{-40}$  °C inside an Ar-filled glove box with water and oxygen content below 0.1 ppm.

To avoid heating during electrode deposition, we make electrical contacts to the exfoliated monolayers by cold-welding indium/gold microelectrodes on the cold stage. Once the device fabrication is complete, we seal each device in a chip carrier (we use ceramic dual-in-line chip carriers) with vacuum grease and a cover glass inside the glove box, and then transfer the whole package into the cryostat. The cover glass comes off once the sample space of the cryostat is evacuated before low-temperature transport measurements, so the doping level can be tuned in situ.

### Fabricating monolayer Bi-2212 for STM/STS measurements

We used a vacuum-compatible tape (Kapton tape with silicone adhesive; Accu-Glass Products) to exfoliate thin flakes of Bi-2212 onto Si wafer (covered with a 285-nm-thick  ${\rm SiO_2}$  layer) in a vacuum chamber with a base pressure of  $1\times 10^{-10}$  mbar. Few-layer Bi-2212 on the substrate exhibits quantized contrast that correlates well with the number of layers (see also Fig. 1). The correlation makes the search (also done in UHV with  $12\times$  Ultrazoom (Navitar) through a re-entrant viewport) for monolayers convenient. Some of the flakes touch electrodes (Cr/Au with thickness of 2 nm and 3 nm, respectively) in the form of stripes that are pre-patterned on the wafer before the exfoliation; we choose these flakes for STM measurements (Fig. 3a). Except for brief moments when the samples were being transferred to the STM stage, the temperature was always kept below -120 °C. Finally, we confirm the thickness of the samples with AFM outside the UHV after all measurements are completed, to ensure that they were indeed monolayers (Extended Data Fig. 6).

# Finite-size scaling analysis of the superconductor-to-insulator transition in monolayer Bi-2212

At the superconductor-to-insulator transition (SIT) in monolayer Bi-2212, HTS emerges from the parent Mott insulator as the sample is doped beyond a critical level. Such a transition is an important example of a continuous quantum phase transition (QPT) that is driven by an external parameter x at absolute zero temperature  $^{59}$ ; the quantum critical point at  $x_c$  separates ground states with different symmetry. Exactly how Cooper pairs form in the 2D copper oxide plane and condense into the superconducting phase is a key outstanding question. However, crucial information on the transition can be obtained by investigating the scaling behaviour of  $R_{\Box}(x, T)$  as x approaches  $x_c$  at finite T. This is accomplished by finite-size scaling analysis under the general scheme

of QPT<sup>60,61</sup>. Near the quantum critical point, the correlation length  $\xi$  and correlation time  $\tau$  become the only characteristic scales in length and time, respectively, and they diverge as  $\xi \propto |x-x_c|^{-\nu a}$  and  $\tau \propto \xi^z \propto |x-x_c|^{-\nu a}$ .

and z are critical exponents. The theory of finite-size scaling asserts that physical quantities have a scaling form that, together with exponents v and z, depend only on global properties of the system, but not on microscopic details. For 2D SIT, the appropriate finite-size scaling form is  $^{59}$ :

$$R_{\Box}(x,T) = R_c f(|x-x_c|T^{-1/\nu z}).$$
 (1)

Here the transition is driven by doping variation, so x = p;  $R_c$  is the critical resistivity at the  $p \rightarrow p_c$  and  $T \rightarrow 0$  limit, and f is a universal scaling function. Such scaling does not depend on the exact value of p, but the exponent vz and the critical resistivity  $R_c$  encode the fundamental properties of the transition. In particular, vz is determined by the universality class that the system belongs to; its value thus provides precious information such as the symmetry of order parameter manifold and types of disorder in 2D Bi-2212 (ref.  $^{48}$ ).

Extended Data Fig. 4a-c illustrates the finite-size scaling analysis of the SIT in sample A. Following the procedure described in ref. 59, we first invert the  $R_{\Box}(p, T)$  data matrix in Extended Data Fig. 4a, and locate the critical point  $p_c$ , where all isotherms converge to  $R_{\Box} = R_c \approx 10.2 \text{ k}\Omega$ (Extended Data Fig. 4b). We then scale the horizontal axis of Extended Data Fig. 4b as u = |p - p|t(T) in Extended Data Fig. 4c. Here a single set of temperature-dependent parameters t(T) can force all curves to collapse to a universal scaling function. Further analysis shows that t(T)follows a power law dependence,  $t(T) \propto T^{-1/1.53}$  (Extended Data Fig. 5a, blue circles). The SIT in monolayer Bi-2212 is, therefore, well described by continuous 2D QPT, with vz = 1.53 matching the critical exponents of the SITs driven by ionic gating in thin films of La<sub>2-x</sub>Sr<sub>x</sub>CuO<sub>4</sub> (LSCO, ref. <sup>48</sup>), lithium-intercalated  $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$  ( $\text{Li}_x\text{Bi-2212}$ , ref. <sup>62</sup>), and  $\text{La}_2\text{CuO}_{4+\delta}$ (LCO, ref. 63). The close match indicates that the SIT transitions in these copper oxides all belong to the same universality class, even though the critical resistivities differ among these systems.

A survey of critical exponents in copper oxide superconductors, however, shows that not all vz agree with the value in monolayer Bi-2212; various vz values were found to cluster around two different values: 3/2 and 7/3 (refs.  $^{48,49,62-64}$ ; Extended Data Fig. 4b, blue squares). It therefore appears that the transitions fall into two distinct universality classes, even though in all copper oxide superconductors the superconductivity arises from doping Mott insulating  $CuO_2$  planes. These observations raise two fundamental questions: (1) what specifically causes the disparate critical exponents in copper oxide superconductors? and (2) what universality classes do they correspond to? We now address these questions by investigating the SIT in monolayer Bi-2212 along another dimension in the parameter space—the disorder level. Here we tune the disorder level by introducing a small amount of air (that contains water vapour, the main degradation agent) into the sample chamber while annealing monolayer Bi-2212 at elevated temperatures.

Extended Data Fig. 4g displays the temperature-dependent resistivity,  $R_{\square}(T)$ , of a monolayer Bi-2212 (sample C). The sample undergoes a sequence of annealing cycles in 10 mbar of air (containing about 0.3 mbar of water vapour) at room temperature. The curves were obtained between each annealing cycle. We observe that the resistivity drops to zero in two steps as the temperature is lowered. The higher-temperature drop occurs at the apparent  $T_{\rm c}$  of the monolayer, but the resistivity drops to zero only after a second transition at a lower temperature (Extended Data Fig. 4g). Such a two-step transition resembles the superconducting transition in 2D Josephson-coupled superconducting arrays 65.66 and is ubiquitous in disordered 2D superconducting systems in general 67.68. A simplified picture captures the basic physics of the two-step transition: the disordered 2D superconductor can be modelled as superconducting islands embedded in normal metal that provide weak Josephson

coupling between the islands. The higher-temperature transition corresponds to the superconducting transition within the islands, but the entire sample becomes superconducting only when the global, interisland phase coherence is established after a second transition at a lower temperature<sup>65</sup>.

The SIT takes place at the lower-temperature transition in this disordered monolayer Bi-2212. Because the apparent  $T_c$  does not change appreciably during the SIT transition (Extended Data Fig. 4g), the transition is now predominantly driven by disorder that mainly affects the metallic region between the islands. Finally, we perform finite-size scaling analysis of the disorder-driven SIT in monolayer Bi-2212. We parameterize the phenomenological disorder level as  $d = \text{const.}/R_{\square}(T = 200 \text{ K})$ . (The value of the constant does not affect our analysis; we chose const.= 213  $\Omega$ .) Using the scaling form (1) with  $x \equiv d$ , we obtained a critical exponent of vz = 2.35 which is close to 7/3. The same analysis on a less disordered monolayer yields a similar vz (Extended Data Fig. 4d and Extended Data Fig. 5a).

We can now explain the two disparate critical exponents observed in copper oxide superconductors. We first note that the two distinct critical exponents in monolayer Bi-2212 confirm early observations that SITs in copper oxide superconductors fall into two universality classes. The mystery is, however, resolved—our results show that the two universality classes characterize the doping-driven SIT in the clean limit and the disorder-driven SIT in the dirty limit, respectively. The exponent vz=7/3 points towards a quantum percolation model that indeed describes a strongly disordered superconductor <sup>69</sup>. Meanwhile, vz=3/2 encodes the essential physics of an intrinsic copper oxide superconductor in both bulk and 2D limits. The fact that bulk and monolayer Bi-2212 belong to the same universality class suggests that the antiferromagnetic order found in bulk Bi-2212 may persist in the monolayer. The microscopic origin of vz=3/2 however, remains an open question that requires further investigation.

### **Data availability**

The datasets generated and analysed during the current study are available from the corresponding author on reasonable request.

- Masteika, V., Kowal, J., Braithwaite, N. S. J. & Rogers, T. A review of hydrophilic silicon wafer bonding. ECS J. Solid State Sci. Technol. 3, Q42–Q54 (2014).
- Sterpetti, E., Biscaras, J., Erb, A. & Shukla, A. Comprehensive phase diagram of twodimensional space charge doped Bi<sub>2</sub>Sr<sub>2</sub>CaCu<sub>2</sub>O<sub>8+t</sub>. Nat. Commun. 8, 2060 (2017).
- Zhao, S. Y. F. et al. Sign reversing Hall effect in atomically thin high temperature superconductors. Phys. Rev. Lett. 122, 247001 (2019).
- Jin, S.-G., Zhu, Z.-Z., Liu, L.-M. & Huang, Y.-L. Water reactions of superconducting Bi<sub>2</sub>Sr<sub>2</sub>CaCu<sub>2</sub>O<sub>8</sub> phase at 0°C and ambient temperature. Solid State Commun. 74, 1087-1090 (1990).
- Gao, W. & Vander Sande, J. B. The degradation behavior of high-T<sub>c</sub> BSCCO/Ag superconducting microcomposites in water. *Mater. Lett.* 12, 47–53 (1991).
- Yun, S. H. & Karlsson, U. O. Water degradation of a- and c-axis oriented HgBa<sub>2</sub>CaCu<sub>2</sub>O<sub>x</sub> superconducting thin films. J. Appl. Phys. 82, 6348 (1997).
- Marković, N., Christiansen, C., Mack, A. M., Huber, W. H. & Goldman, A. M. Superconductorinsulator transition in two dimensions. *Phys. Rev. B* 60, 4320–4328 (1999).
- Fisher, M. P. A. Quantum phase transitions in disordered two-dimensional superconductors. *Phys. Rev. Lett.* 65, 923–926 (1990).

- Sondhi, S. L., Girvin, S. M., Carini, J. P. & Shahar, D. Continuous quantum phase transitions. Rev. Mod. Phys. 69, 315–333 (1997).
- Liao, M. et al. Superconductor-insulator transitions in exfoliated Bi<sub>2</sub>Sr<sub>2</sub>CaCu<sub>2</sub>O<sub>8+5</sub> flakes. Nano Lett. 18, 5660–5665 (2018).
- Garcia-Barriocanal, J. et al. Electronically driven superconductor-insulator transition in electrostatically doped La<sub>2</sub>CuO<sub>4+6</sub> thin films. *Phys. Rev. B* 87, 024509 (2013).
- Zeng, S. W. et al. Two-dimensional superconductor-insulator quantum phase transitions in an electron-doped cuprate. Phys. Rev. B 92, 020503 (2015).
- Eley, S., Gopalakrishnan, S., Goldbart, P. M. & Mason, N. Approaching zero-temperature metallic states in mesoscopic superconductor-normal-superconductor arrays. *Nat. Phys.* 8, 59–62 (2012).
- Bøttcher, C. G. L. et al. Superconducting, insulating and anomalous metallic regimes in a gated two-dimensional semiconductor-superconductor array. *Nat. Phys.* 14, 1138–1144 (2018).
- Allain, A., Han, Z. & Bouchiat, V. Electrical control of the superconducting-to-insulating transition in graphene-metal hybrids. *Nat. Mater.* 11, 590–594 (2012).
- Chen, Z. et al. Carrier density and disorder tuned superconductor-metal transition in a two-dimensional electron system. Nat. Commun. 9, 4008 (2018).
- Steiner, M. A., Breznay, N. P. & Kapitulnik, A. Approach to a superconductor-to-Boseinsulator transition in disordered films. Phys. Rev. B 77, 212501 (2008).
- Kim, D. H., Goldman, A. M., Kang, J. H. & Kampwirth, R. T. Kosterlitz-Thouless transition in Tl,Ba,CaCu,O<sub>8</sub> thin films. *Phys. Rev. B* 40, 8834–8839 (1989).
- Aslamasov, L. G. & Larkin, A. I. The influence of fluctuation pairing of electrons on the conductivity of normal metal. *Phys. Lett. A* 26, 238–239 (1968).
- Ito, T., Takenaka, K. & Uchida, S. Systematic deviation from T-linear behavior in the inplane resistivity of YBa<sub>2</sub>Cu<sub>3</sub>O<sub>7-y</sub>: evidence for dominant spin scattering. Phys. Rev. Lett. 70, 3995–3998 (1993).
- Hüfner, S., Hossain, M. A., Damascelli, A. & Sawatzky, G. A. Two gaps make a hightemperature superconductor? Rep. Prog. Phys. 71, 062501 (2008).

Acknowledgements We thank D.-H. Lee, Z.-Y. Weng, Q.-K. Xue, S.-W. Cheong, Y. Wang, H. Ding and H. Luo for discussions. We also thank X. Jin and D. Feng for their help with the experiment. Part of the sample fabrication was conducted at Nano-fabrication Laboratory at Fudan University, Work at Brookhayen National Laboratory was supported by the Office of Science. US Department of Energy under contract no. DE-SC0012704. Y.Y., L.M. and Y.Z. acknowledge support from the National Key Research Program of China (grant nos. 2016YFA0300703, 2018YFA0305600), National Science Foundation of China (grant nos. U1732274, 11527805, 11425415 and 11421404), Shanghai Municipal Science and Technology Commission (grant no. 18JC1410300) and Strategic Priority Research Program of Chinese Academy of Sciences (grant no. XDB30000000). Y.Y. acknowledges support from the National Postdoctoral Program for Innovative Talents (grant no. BX20180076) and China Postdoctoral Science Foundation (grant no. 2018M641907). P.C. acknowledges support from National Postdoctoral Program for Innovative Talents (grant no. BX201600036), Shanghai Sailing Program (grant no. 17YF1429000), Shanghai Municipal Natural Science Foundation (grant no. 17ZR1442400) and China Postdoctoral Science Foundation (grant no. 2017M610221). X.H.C. acknowledges support from the National Science Foundation of China (grant no. 11888101, 11534010), the National Key R&D Program of China (grant no. 2017YFA0303001 and 2016YFA0300201), Strategic Priority Research Program of the Chinese Academy of Sciences (grant no. XDB25000000) and the Key Research Program of Frontier Sciences, CAS (grant no. QYZDY-

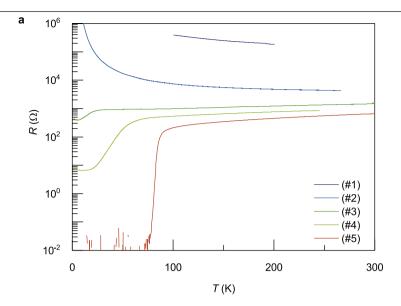
**Author contributions** The order of the first two authors was determined arbitrarily. Y.Z. conceived the project. Y.Z. and X.H.C. supervised the experiments. R.Z. and G.D.G. synthesized bulk crystals. Y.Y., L.M. and C.Y. developed sample fabrication techniques. Y.Y. did transport measurements. L.M. led the STM study. L.M., P.C. and C.Y. did STM measurements and J.S. provided support. Y.Y., L.M., P.C. and Y.Z. analysed the data and wrote the paper with input from all authors.

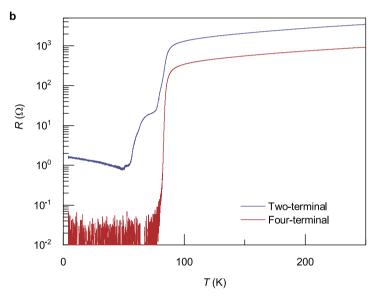
Competing interests The authors declare no competing interests.

### Additional information

Correspondence and requests for materials should be addressed to L.M., X.H.C. or Y.Z. Peer review information Nature thanks Tetsuro Hanaguri and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

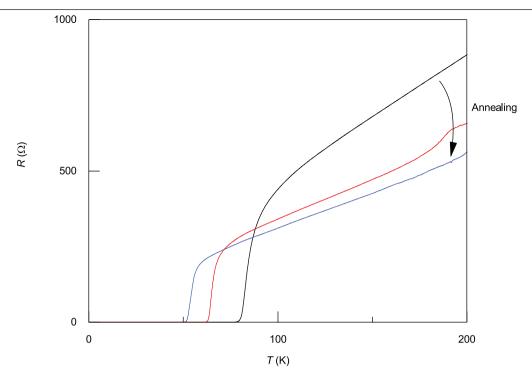
Reprints and permissions information is available at http://www.nature.com/reprints.





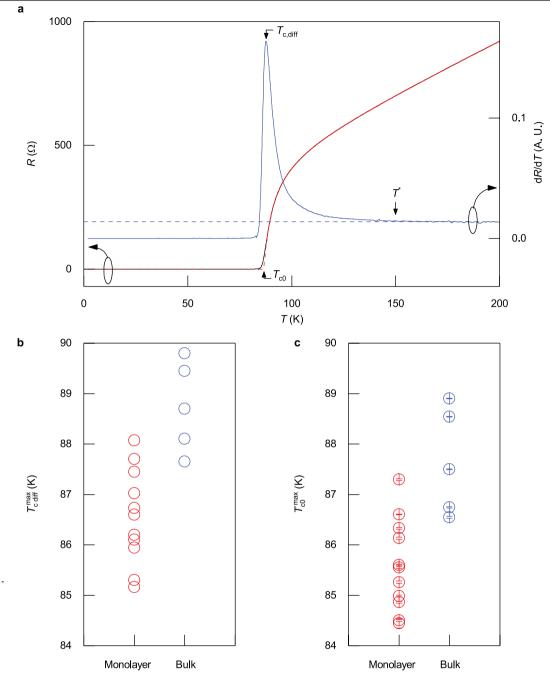
 $\label{lem:continuous} \textbf{Extended Data Fig. 1} | \textbf{Transport properties of typical monolayer Bi-2212} \\ \textbf{samples fabricated by various methods. a}, \textbf{Temperature-dependent} \\ \textbf{resistance of monolayer Bi-2212 samples. Here (#1)-(#5) refer to five typical samples fabricated by different methods indicated in Extended Data Table 1.} \\ \textbf{b}, \textbf{Resistance of a typical cold-welded Bi-2212 monolayer device measured with} \\ \end{aligned}$ 

two-terminal (blue) and four-terminal (red) configurations. The four-terminal configuration is adopted in all our measurements presented in the main text, because it eliminates spurious signals from electrical contacts. The two-terminal resistance in the superconducting state gives an estimate of the contact resistance of the order of  $1\Omega$ .



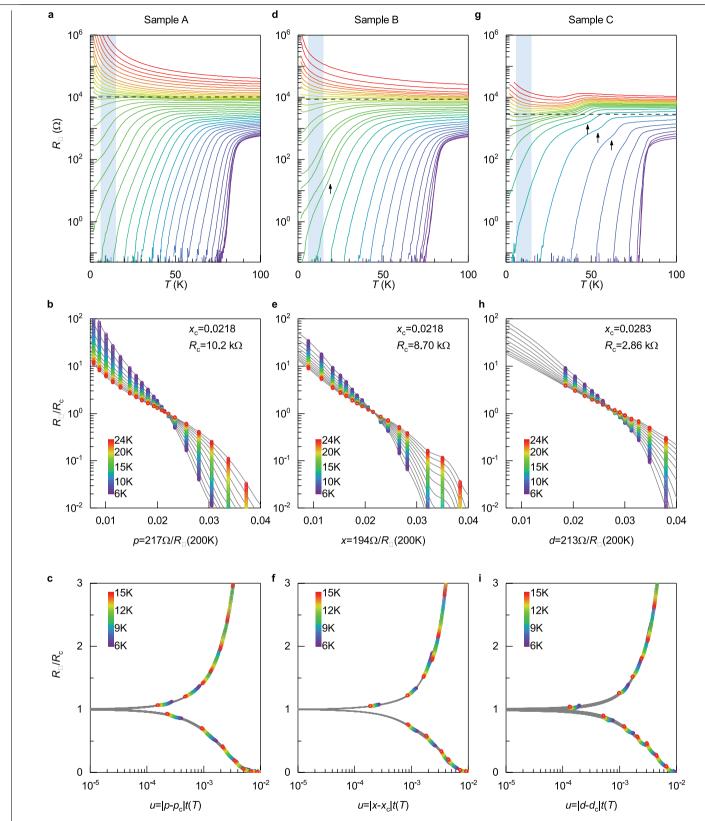
Extended Data Fig. 2 | Temperature-dependent resistance of a monolayer Bi-2212 sample annealed in ozone. Annealing cycles were performed under an  $O_3$  partial pressure of about 50 Pa at temperatures between 220 K and 240 K.  $O_3$  was purged with helium gas between annealing cycles, and data were obtained in helium vapour. Each annealing cycle lasts 5–30 min. Monolayer

Bi-2212 was initially at optimal doping (black curve). The annealing cycles progressively increase the doping level of the sample. The red curve was obtained after first annealing, and blue curve was obtained after second annealing.



**Extended Data Fig. 3** | **Extracting**  $T_c$  **and**  $T^*$  **from temperature-dependent resistance of monolayer Bi-2212. a**, Illustration of  $T_c$  and  $T^*$  extraction from temperature-dependent resistance (black curve, which mostly overlaps with the red curve) and its derivative (blue curve). We used two definitions of  $T_c$  in our analysis: (i)  $T_{c,diff}$  where the slope of resistance vs temperature curve is maximum<sup>70</sup>; (ii)  $T_{c,0}$  from fitting with Aslamasov–Larkin paraconductivity model<sup>71</sup>  $\Delta \sigma = \sigma(T) - \sigma_{normal}(T) = a(T/T_{c,0} - 1)^{-1}$ . Near optimal doping,  $\sigma_{normal}(T) = (bT + c)^{-1}$ , so  $T_{c,0}$  can be extracted from fitting with

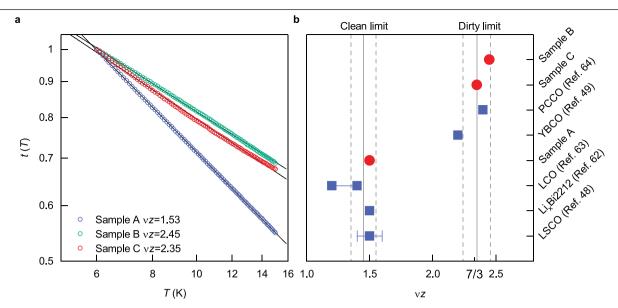
 $R(T) = (bT+c)(T-T_{c0})/(T-T_{c0}+a) \ (\text{red curve}). \ T^* \ is \ determined as \ the temperature at which the derivative of temperature-dependent resistance deviates from constant value (broken blue line; ref. \ ^2\).$ **b, c,**  $<math>T_{c, \text{diff}}^{\text{max}}(\mathbf{c})$  of monolayer and bulk Bi-2212. Bulk data were obtained from optimally doped crystals (OP88). Under both definitions, the highest maximum  $T_c$  of monolayers is within the statistical uncertainty range of the  $T_c$  in optimally doped bulk crystals.



**Extended Data Fig. 4** | See next page for caption.

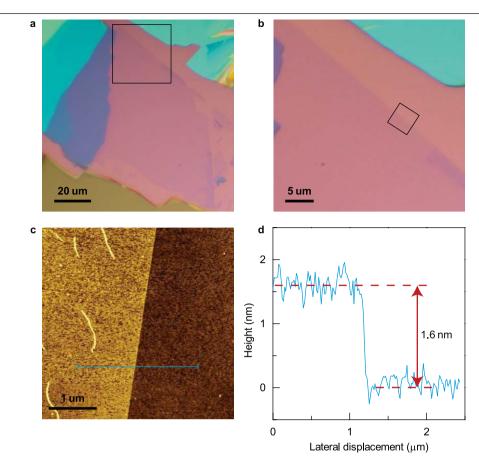
Extended Data Fig. 4 | Superconductor-insulator transition in monolayer **Bi-2212.** a, Temperature-dependent resistivity  $R_{\square}(p,T)$  of sample A. The doping level, fixed for each curve, is tuned by repeated annealing cycles under vacuum (pressure below 10<sup>-4</sup> mbar). The initially superconducting sample becomes insulating via a QPT. Broken line marks the separatrix where the transition occurs. Blue shaded region indicates the temperature range in which we perform the finite-size scaling analysis; the slight up-turn in resistivity at lower temperatures suggests intermediate phase or additional QCP between the superconducting and insulating phases<sup>49</sup>. **b**, Same dataset in **a** plotted inversely, that is,  $R_{\Box}(p, T)$ plotted as a function of doping level at fixed temperatures between 6 K and 24 K. Each colour refers to a fixed temperature. Continuous curves are interpolations of data points at different temperatures. The point where all curves cross defines the critical point the QPT, ( $R_c = 10.2 \pm 0.6 \text{ k}\Omega$ ,  $p_c = 0.022 \pm 0.002$ ). **c**, Scaling of the same data with respect to variable  $u = p - p_c | \tilde{t}(T)$ . A single set of temperaturedependent parameters t(T) can force all data to collapse to a universal scaling function on both sides of the SIT. d, Temperature-dependent resistivity of sample B. Data were obtained between annealing cycles performed under  $10^{-1}$  mbar of air that contains about  $3 \times 10^{-3}$  mbar of water vapour. The annealing cycles progressively increase the normal state resistivity, and induces SIT in the monolayer. Blue shaded region marks the temperature range in which we perform the finite-size scaling analysis.e, Same resistivity data in d plotted as a

function of  $x = 194 \Omega/R_{\Box}(T = 200 \text{ K})$ . Here x is a phenomenological variable that parametrizes the external factor (doping or disorder level) that drives the SIT; the precise value of x does not affect the finite-size scaling analysis according to formula (1). The critical point of the SIT is identified as  $(R_c = 8.7 \pm 0.6 \text{ k}\Omega, x_c = 0.022 \pm 0.002)$ . **f**, Scaling analysis of the dataset in **e**. The analysis yields a critical exponent of vz = 2.45. The vz differs from the critical exponent in doping-driven SIT in sample A, but coincides with the value in disorder-driven SIT in sample C. Similar to sample C, sample B also features a two-step superconducting transition (marked by black arrow) that indicates considerable amount of disorder. We therefore conclude that disorder level  $drives \ the \ SIT \ in \ sample \ B. \ \textbf{g}, Temperature-dependent \ resistivity \ of \ sample \ C.$ Curves are obtained between annealing cycles performed under about 10 mbar of air. Such annealing cycles introduce disorders into the monolayer, and the superconductivity transition occurs in two steps. The disorder-driven SIT takes place at the lower-temperature transition (blue shaded region). h, Inverse of the dataset in g. Horizontal axis represents the phenomenological disorder level that is parametrized as  $d = 213 \,\Omega/R_{\Box}(T = 200 \,\mathrm{K})$ . Smooth interpolations of the data points cross at the critical point ( $R_c = 2.86 \pm 0.17 \text{ k}\Omega$ ,  $x_c = 0.028 \pm 0.002$ ). i, Scaling of the same data in h with respect to variable  $u = |d - d_c|t(T)$ . t(T) is chosen such that all data collapse to a universal scaling function.



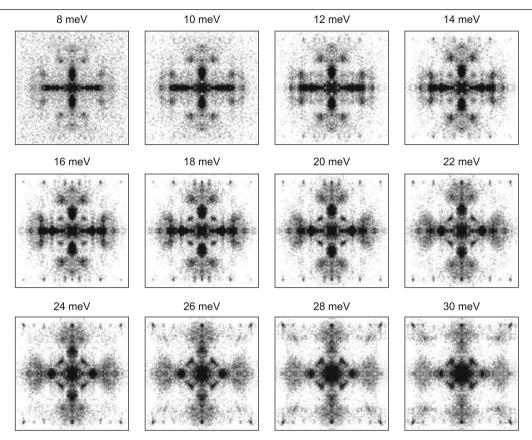
**Extended Data Fig. 5 | Critical exponents of superconductor-insulator transitions in copper oxide superconductors. a**, Temperature-dependent parameter t(T) obtained from finite-size scaling analysis in Extended Data Fig. 4. Values of t(T) from all three monolayer Bi-2212 samples follow power-law dependence; the slope of the line fits (solid lines) yields the critical exponents of the SIT vz = 1.53, 2.45 and 2.35 for samples A, B and C, respectively. **b**, Critical

exponents vz obtained in monolayer Bi-2212 (red circles) and various other copper oxide superconductors (black squares). All vz fall into the neighbourhood of one of the two values, 3/2 and 7/3, that characterize the SIT in the clean and dirty limit, respectively (see text). Solid vertical lines mark the mean, and broken lines the standard deviation, of the vz values in each category.



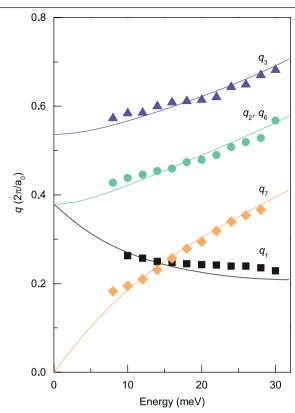
Extended Data Fig. 6 | Characterization of monolayer Bi-2212 after STM measurements. a, Optical image of typical Bi-2212 flakes exfoliated on SiO $_2$ /Si substrate. The monolayer (light purple region in the centre) is identified from its optical contrast. b, A magnified view of the area marked by the square in a. c, AFM topography of the area marked by the square in b. Both the optical image

and the AFM topography were obtained in an Ar atmosphere inside a glove box after STM measurements performed in UHV.  $\mathbf{d}$ , Line cut of the AFM topography along the line shown in  $\mathbf{c}$ . The step height of about 1.6 nm confirms that the Bi-2212 flake measured in STM was indeed a monolayer.

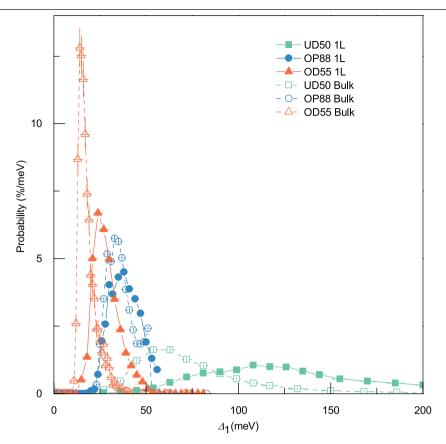


Extended Data Fig. 7 | Fourier transform of the conductance ratio map obtained on monolayer Bi2212 at various energies. Each panel displays a Fourier transform of the conductance ratio map  $Z(\mathbf{r},E)$  of nearly optimally doped monolayer Bi-2212 at the energy labelled on the panel. The  $Z(\mathbf{r},E)$  maps

are obtained from a set of  $200\times200$ -pixel conductance maps taken on an area of  $500\,\text{Å}\times500\,\text{Å}$  with an energy resolution of  $2\,\text{meV}$ . Data were obtained from the same sample in Fig. 4 (here we show the full dataset).

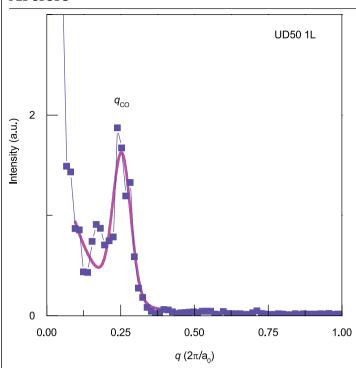


**Extended Data Fig. 8** | **Energy dispersion of the q-vectors.** Amplitudes of measured  $\mathbf{q}_i$  (in units of  $2\pi/a_0$ ) are plotted as functions of energy (i=1...7, except that  $\mathbf{q}_4$  and  $\mathbf{q}_5$  are too weak to be detected). We followed the method described in ref. <sup>23</sup> to obtain  $\mathbf{q}_i$ . Solid lines are energy dispersion of the  $\mathbf{q}$ -vectors expected in the octet model.

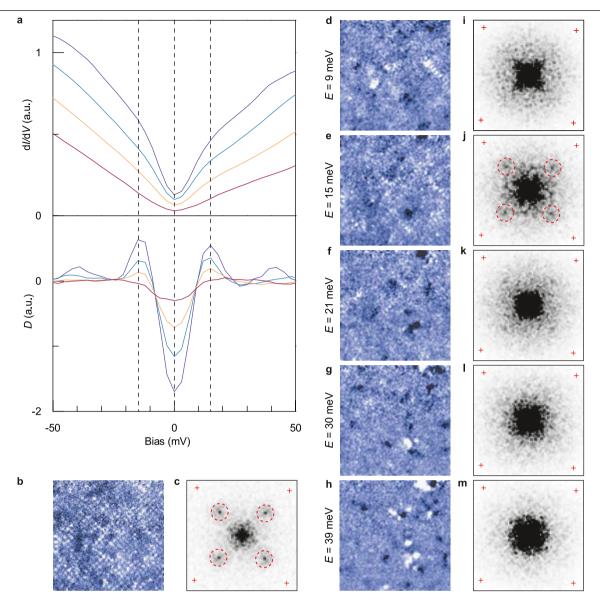


**Extended Data Fig. 9** | **Histograms of**  $\Delta_1$ ( $\mathbf{r}$ ) gap maps in monolayer and bulk **Bi-2212.** Solid and empty symbols represent data from monolayer and bulk Bi-2212, respectively.  $\Delta_1$  distributions in monolayers shift towards higher energies compared with those in bulk crystals. The shift reflects slight loss of oxygen doping during monolayer sample fabrication. Specifically, the doping level p is directly related to the average value of the pseudogap. From the average pseudogap, we estimate that p = 0.06±0.02, 0.16±0.02 and 0.19±0.02 for

monolayers obtained from UD50, OP88 and OD55, respectively  $^{23,36,73}$ . These values are lower than the doping levels extracted in the bulk crystals  $(p=0.08\pm0.02,\ 0.17\pm0.02$  and  $0.22\pm0.01$  for UD50, OP88 and OD55, respectively). Here we used the relations  $2\Delta_1=152\ \text{meV}\times(0.27-p)/0.22$  for 0.1< p<0.22 and  $2\Delta_1=85\ \text{meV}\times(0.12-p)/0.02$  for 0.06< p<0.08 to estimate the doping level in both bulk crystals and monolayers.



**Extended Data Fig. 10** | **Wavevector of the CDW order in monolayer Bi-2212 obtained in UD50.** Line cut (blue line) of the FFT of g ( $\mathbf{r}, E$  = 20 meV) map in Fig. 5h along the Cu–O bond direction exhibits a peak at  $q_{\rm CO}$  = 0.25 ( $2\pi/a_0$ ) that is associated with the charge-ordered state. The magenta line is a Gaussian fit to the peak plus a decaying exponential background. The full-width at half-maximum of the peak yields a correlation length of about  $14a_0$ .



Extended Data Fig. 11 | Pair density wave in monolayer Bi-2212. a, Four representative conductance spectra (dI/dV; upper panel) and the negative of their second derivative ( $D=-d^3I/dV^3$ ; lower panel) in under-doped monolayer Bi-2212 obtained from UD50. We additionally define  $H=dI/dV(E=\Delta_0)-dI/dV(E=0)$ , which corresponds to the amount of low-energy DOS gapped out by Cooper pairing (here  $\Delta_0=15$  meV). The pair density wave can be visualized by spatially mapping either H or D (ref.  $^{32}$ ). b, H(r) map on a 40 nm × 40 nm area. A chequerboard pattern is clearly resolved. c, Fourier transform of the H(r) map in b. Peaks at  $|\mathbf{q}|=(0.25\pm0.02)2\pi/a_0$  (marked by broken circles) along the Cu–O bond directions indicate the emergence of pair density wave order  $^{32}$ . d–h, D(r) maps obtained on the same area in b at various energies. i–m, Fourier transform

of the  $D(\mathbf{r})$  maps in  $\mathbf{d}$ - $\mathbf{h}$ . The  $|\mathbf{q}|=2\pi/4a_0$  spatial modulations at E=15 meV (broken circles in  $\mathbf{j}$ ) again indicate the existence of pair density wave<sup>32</sup>. Red crosses mark  $\mathbf{q}=(0,\pm\pi/a_0)$  and  $(\pm\pi/a_0,0)$ . We followed the method described in ref.  $^{32}$  to obtain  $H(\mathbf{r})$  and  $D(\mathbf{r})$  maps. First, a set of conductance (dl/dV) spectra was taken on a  $160\times160$  grid over the 40 nm  $\times40$  nm area. Here we used a set-point bias voltage of -300 mV, which is far beyond the energy scale of the charge-ordered state, to eliminate possible set-point effects. We then fitted each dl/dV spectrum with a second-order polynomial, and took the second derivative of the polynomial to obtain the D spectrum. The  $H(\mathbf{r})$  map is directly obtained from the dl/dV spectra grid.

# Extended Data Table 1 | Optimizing fabrication process for monolayer and bilayer Bi-2212 samples

	Contact method	Air- exposure time	Total fabrication time in glove box	Bulk crystal	# of devices	T <sub>c</sub>
Bilayer	Metal evaporation at room temperature	5 min	4 h	OP88	5	50 - 70 K
	Pre-patterned bottom contact	2 s	2 h	OP88	1	~ 70 K
	Pre-patterned bottom contact	2 s	2 h	OD55	1	80 - 90 K
	Cold welding	/	2 h	OP88	4	80 - 90 K
Monolayer	Metal evaporation at room temperature	5 min	4 h	OP88	2	Insulating
	Prepatterned bottom contact <sup>(#1)</sup>	2 s	2 h	OP88	3	Insulating
	Pre-patterned bottom contact	2 s	2 h	OD55	1	Insulating
	Metal evaporation at room temperature <sup>(#2)</sup>	/	4h	OP88	2	Insulating
	Metal evaporation at low temperature (~ 100 K) <sup>(#3)</sup>	/	4 h	OD55	3	< 10 K
	Cold welding <sup>(#4)</sup>	/	2 h	OP88	2	< 40 K
	Cold welding	/	2 h	OD55	4	70 - 80 K
	Cold welding <sup>(#5)</sup>	/	0.5 - 1 h	OD55	14	80 - 90 K

We have systematically investigated the effects of the following key factors on the transport properties of monolayer and bilayer Bi-2212: contact method, air-exposure time and total fabrication time in glove box. We observe that monolayer Bi-2212 is more prone to degradation than is bilayer graphene. In particular, exposure to air is most detrimental to sample quality of the monolayers. Evaporating metal contacts (through shadow mask) also causes considerable degradation. We find that cold-welding indium contacts in the glove box preserves the monolayer sample quality. Here, thin indium foils make stable contacts with a thick flake that is connected to the monolayer, so that the thick flake electrically bridges the indium electrodes and the monolayer (Fig. 1e). The monolayer samples exfoliated from an over-doped crystal (OD55) are slightly over-doped, and their maximum  $T_c$  is comparable to the  $T_c$  of optimally doped bulk crystals (Fig. 2c).

(\*\*I)-(\*\*E)</sup>Transport properties of typical monolayer samples from these categories are shown in Extended Data Fig. 1.

# Extended Data Table 2 | Annealing sequence of monolayer Bi-2212

Annealing temperature	Annealing time	Doping regime	$\Delta_1$ (meV)
As-exfoliated	/	Over-doped	28 ± 1
25 ℃	1 week	Nearly optimally doped	40 ± 1
130 °C	30 min	Under-doped	56 ± 1
220 °C	30 min	Under-doped	122 ± 4
265 °C	30 min	Extremely under-doped	250 ± 20

This sequence relates to the monolayer Bi-2212 shown in Fig. 6. The as-exfoliated monolayer Bi-2212 (over-doped;  $\Delta_1$  = 28±1 meV) was annealed under UHV with a base pressure of 1×10<sup>-10</sup> mbar. The pseudogaps  $\Delta_1$  were extracted from spatially averaged conductance spectra.

# Nanomagnetic encoding of shape-morphing micromachines

https://doi.org/10.1038/s41586-019-1713-2

Received: 25 April 2019

Accepted: 4 September 2019

Published online: 6 November 2019

Jizhai Cui<sup>1,2,4\*</sup>, Tian-Yun Huang<sup>3,4\*</sup>, Zhaochu Luo<sup>1,2</sup>, Paolo Testa<sup>1,2</sup>, Hongri Gu<sup>3</sup>, Xiang-Zhong Chen<sup>3</sup>, Bradley J. Nelson<sup>3</sup> & Laura J. Heyderman<sup>1,2</sup>

Shape-morphing systems, which can perform complex tasks through morphological transformations, are of great interest for future applications in minimally invasive  $medicine^{1.2}$ ,  $soft robotics^{3-6}$ ,  $active metamaterials^7$  and  $smart surfaces^8$ . With current fabrication methods, shape-morphing configurations have been embedded into structural design by, for example, spatial distribution of heterogeneous materials<sup>9-14</sup>, which cannot be altered once fabricated. The systems are therefore restricted to a single type of transformation that is predetermined by their geometry. Here we develop a strategy to encode multiple shape-morphing instructions into a micromachine by programming the magnetic configurations of arrays of singledomain nanomagnets on connected panels. This programming is achieved by applying a specific sequence of magnetic fields to nanomagnets with suitably tailored switching fields, and results in specific shape transformations of the customized micromachines under an applied magnetic field. Using this concept, we have built an assembly of modular units that can be programmed to morph into letters of the alphabet, and we have constructed a microscale 'bird' capable of complex behaviours, including 'flapping', 'hovering', 'turning' and 'side-slipping'. This establishes a route for the creation of future intelligent microsystems that are reconfigurable and reprogrammable in situ, and that can therefore adapt to complex situations.

It has been a long-standing goal to create intelligent machines that are untethered and can execute tasks at small scales. Magnetic actuation is of particular interest for the control of these machines, because it comes with the advantage of being able to perform tasks in confined and enclosed spaces<sup>15</sup>. In magnetic shape-morphing systems, a mechanical torque is generated when the magnetization of a magnetic medium is not in line with the applied magnetic field<sup>16</sup>. Previous programming of the magnetic configurations could be achieved by reorienting permanent-magnet microparticles in millimetre-sized devices 10,17,18. For micrometre-sized devices, programming methods such as aligning superparamagnetic nanoparticles 12,19 (region i in Fig. 1a) and selective coating with soft magnetic thin films  $^{20}$  (region iii in Fig. 1a) have been demonstrated. In this work, we used stadium-shaped singledomain nanomagnets to encode shape-morphing information into micromachines. The use of nanomagnets with lateral dimensions in an intermediate range, between 100 nm and 500 nm (region ii in Fig. 1a), means they are single domain with a stable remanent magnetization and have a tunable magnetic anisotropy at room temperature<sup>21,22</sup>. As a result of the magnetic shape anisotropy, the magnetization is parallel to the long axis of the magnets, pointing in one of the two directions. By implementing arrays of these nanomagnets in a micromachine, the magnetic configuration can be remotely programmed by applying a sequence of magnetizing fields to store the shape-morphing information.

Inspired by origami<sup>23</sup>, the art of paper folding, the micromachine is designed with two types of component: rigid panels, some of which are made functional with arrays of single-domain nanomagnets covering the panel surface; and structured 'soft' spring hinges as the connecting creases. In the micromachine design shown in Fig. 1b, there are four panels patterned with 60-nm-thick nanomagnets that can be remotely encoded and manipulated, and one central passive panel (see Extended Data Fig. 1 for a scanning electron microscopy, SEM, image). Arrays of nanomagnets with different magnetic switching fields are fabricated on opposite panels—for example, panels I and II in Fig. 1b. The switching fields of the nanomagnets were engineered by varying the aspect ratios of the nanomagnets while maintaining the same volume (see Methods section 'Nanomagnet design and coercivity' and Extended Data Fig. 2). As given by the magnetic hysteresis loops in Fig. 1c (see Methods section 'Magnetic characterization and encoding'), the coercive fields  $B_c$  required to switch the magnets range from about 30 mT for low-aspect-ratio 'wide' nanomagnets (300 nm × 110 nm, nanomagnet type IV), to about 140 mT for high-aspect-ratio 'narrow' nanomagnets (520 nm × 60 nm, nanomagnet type I). The square shape of the loops indicates that they are fully magnetized at remanence. Arrays of type I and type II nanomagnets are patterned on the opposite panels of the four-panel micromachine, and have switching fields of  $B_c(I) \approx 140 \text{ mT}$ and  $B_c(II) \approx 90$  mT, respectively. Type III and type IV nanomagnets with  $B_c(III) \approx 70 \text{ mT}$  and  $B_c(IV) \approx 30 \text{ mT}$  are employed later in this work.

<sup>1</sup>Laboratory for Mesoscopic Systems, Department of Materials, ETH Zurich, Zurich, Switzerland. <sup>2</sup>Laboratory for Multiscale Materials Experiments, Paul Scherrer Institute, Villigen, Switzerland. <sup>3</sup>Institute of Robotics and Intelligent Systems. ETH Zurich, Zurich, Switzerland. <sup>4</sup>These authors contributed equally: Jizhai Cui, Tian-Yun Huang. \*e-mail: huangt@ethz.ch: iizhai.cui@psi.ch

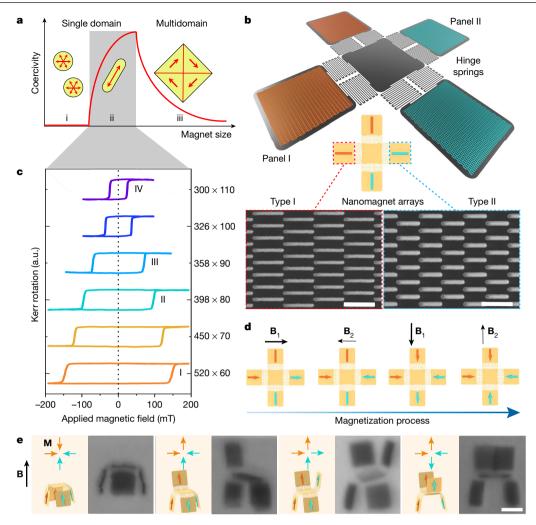


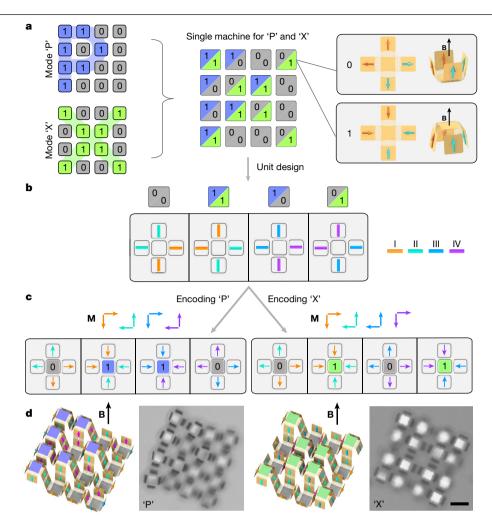
Fig. 1 | Design of a four-panel shape-morphing micromachine. a, Schematic diagram of the magnetic states found in magnets with increasing size: i, superparamagnetic; ii, stable single domain at room temperature; and iii, multidomain state. The red arrows indicate possible magnetization directions. The red trace schematically illustrates the size dependence of the coercivity of the magnets. Nanomagnets in region ii (the shaded area) are implemented in this work. b, Top, four-panel micromachine with an array of 520 nm × 60 nm (type I) nanomagnets on panel I and 398 nm × 80 nm (type II) nanomagnets on panel II; bottom, corresponding SEM images of the nanomagnet arrays. The zig-zag hinge spring has six turns. c, Magneto-optical Kerr effect hysteresis loops of single-domain nanomagnets with the same volume but with six different aspect ratios. The lateral dimensions are indicated to the right of the vertical axis in nanometres. The coercive fields of

the type I–IV nanomagnets are  $B_c(I) \approx 140 \text{ mT}$ ,  $B_c(II) \approx 90 \text{ mT}$ ,  $B_c(III) \approx 70 \text{ mT}$  and  $B_c(IV) \approx 30 \text{ mT.}$  d, Schematic of the encoding of the micromachine using two fields,  $B_1 > B_c(I)$  (large enough to switch type I and type II nanomagnets) and  $B_c(I) > B_2 > B_c(II)$  (large enough to switch type II, but not type I, nanomagnets) applied along both the horizontal and vertical directions (see main text for details). e, Schematics of the magnetic configurations (with type I and type II nanomagnets) and micromachine folding behaviour on application of the controlling magnetic field B = 15 mT, with optical microscope images showing the four different conformations of the fabricated devices. Going from left to  $right, the \, numbers \, of \, panels \, folding \, up/down \, are \, 4/0, 3/1, 2/2 \, (opposite \, panels \, are \, 4/0, 3/1, 2/2) \, (opposite \, panels \, are \, 4/$ having different folding directions) and 2/2 (opposite panels having the same folding direction). Scale bars: 500 nm (b), 10 µm (all images in e).

The micromachine can then be encoded by a series of magnetizing fields, shown schematically in Fig. 1d. By applying a magnetic field  $B_1 > B_c(I)$  along the x direction, both type I and type II nanomagnets are magnetized with full remanent magnetization. A second, lower, magnetic field  $B_c(I) > B_2 > B_c(II)$  is then applied in the opposite direction to remagnetize the type II magnets in the opposite direction, so that the magnetization of the arrays on the two panels point head-tohead along the x direction (orange and turquoise horizontal arrows in Fig. 1d). Similarly, applying  $B_1$  and  $B_2$  fields sequentially along the y direction gives head-to-head magnetization for the other two panels, so that the device has all four panels magnetized towards the centre. Furthermore, using magnetizing field protocols with different combinations of  $B_1$  and  $B_2$  fields, the same micromachine can be magnetized into different magnetic configurations and, since each panel can be magnetized in one of two opposite directions,

there are a total of  $2^4 = 16$  magnetic configurations for the same micromachine.

After programming the magnetic configurations, the micromachine is released from the substrate (see Methods section 'Sample fabrication') and actuated with an applied magnetic field **B** that provides a magnetic torque  $\tau = \mathbf{m} \times \mathbf{B}$  on the panels, where  $\mathbf{m}$  is the total magnetic moment of the nanomagnet arrays on a given panel. All of the panels patterned with nanomagnets try to align with the applied magnetic field direction, which is counterbalanced by the mechanical torque from the deformed hinge springs. (For spring designs and mechanical calculations, see Methods section 'Hinge spring design and properties' and Extended Data Fig. 3) When actuated by the controlling field, the four types of magnetic configuration give four distinct conformations, as shown in Fig. 1e and Supplementary Video 1. The transformations of our micromachines require actuation fields (<15 mT) that are smaller



 $Fig.\,2\,|\,Encoding\,letters\,of\,the\,alphabet\,into\,a\,shape-morphing\,micromachine\,assembled\,from\,an\,array\,of\,4\times4\,four-panel\,units.$ 

**a**, Conceptual design of a micromachine with 16 four-panel units that can be encoded to transform into letters 'P' and 'X', as illustrated by the pair of schematics on the left. Each box in the schematic represents a four-panel unit, where all the magnetizations of the panels in a given unit can be encoded to point outwards or inwards. Consequently, the central panel will move down or up when an out-of-plane controlling field B is applied and we assign these to be the 'O' state or the '1' state, respectively, as illustrated by the insets on the right. In order to combine mode 'P' and mode 'X' into a single micromachine, as illustrated by the central schematic, each unit is assigned one of four possible coding states given by (P, X) = (0, 0), (1, 1), (1, 0) or (0, 1). **b**, Design of the four

different four-panel units with the four different coding states. Each panel contains an array of identical magnets whose orientation is given by the coloured bar on the panels. The colour of the bars corresponds to the colour of the hysteresis loops for the four different types of nanomagnets, I–IV, shown in Fig. 1c. c, Encoding the magnetization in the arrays of nanomagnets of the micromachine for the 'P' and 'X' shape morphing. The magnetization directions are given by the pairs of arrows labelled  $\mathbf{M}$ , and the arrows within the boxes, which are coloured according to the four types of nanomagnet.  $\mathbf{d}$ , Schematics and corresponding optical microscope images of the fabricated devices encoded for 'P' and 'X' shape morphing. The devices are actuated with B=15 mT. Scale bar (for both images), 20  $\mu m$ .

than the switching fields, which are in the range 30–140 mT for the type I–IV nanomagnets. The complete set of shape transformations for all 16 magnetic configurations for this four-panel micromachine are shown in Extended Data Fig. 4. There are four distinct conformations (indicated by the four different background colours) due to the four-fold rotational symmetry of this particular micromachine. Nevertheless, the full programmability is given by the 16 different magnetic configurations, which can give additional conformations when assembling the micromachines into multicomponent devices or when using an asymmetric machine design. We have therefore demonstrated that, by encoding magnetic configurations through magnetizing field protocols, micromachines having the same structural design can be programmed with different shape-morphing behaviours.

Multicomponent shape-morphing micromachines can be constructed by assembling modular units such as the four-panel devices shown in Fig. 1e. We first built a micromachine by assembling the same modular units into a 3 × 3 array, which provides four distinct conformations (see Extended Data Fig. 5 and Supplementary Video 2). Furthermore, multiple tailored conformations can be attained by customizing the individual units and encoding their magnetic configurations. To demonstrate this, we engineered a micromachine that can transform to two distinct letters of the alphabet, 'P' and 'X', using different nanomagnetic encoding, as shown in Fig. 2. For this, we first selected a 4 × 4 assembly of the four-panel units (Fig. 2a) and, when applying the controlling magnetic field, each of the units moves either 'up' (magnetization on all four panels points inwards, represented by a '1' state), or 'down' (magnetization on all four panels points outwards, represented by a '0' state), with the set of '1'-state units representing a letter. Building both mode 'P' and mode 'X' into a single machine gives four distinct coding states -(P, X) = (0, 0), (1, 1), (1, 0)or (0,1)—for each unit, where the first digit is associated with the 'P' mode and the second digit is associated with the 'X' mode. We created four types of assembly unit, each having one of the four coding states, which

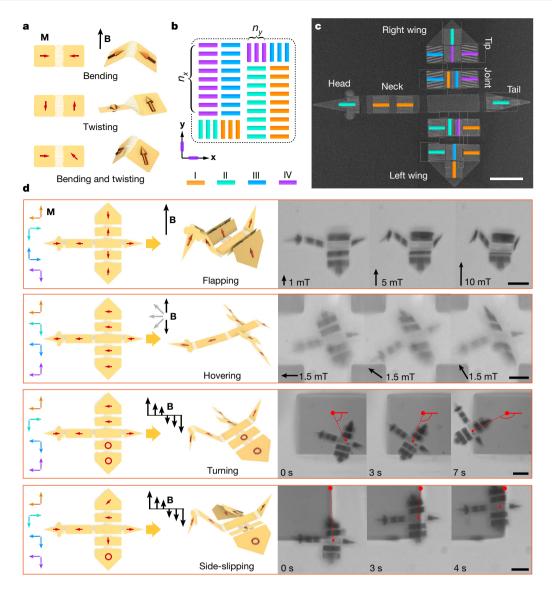


Fig. 3 | Origami-like microscale 'bird' with multiple shape-morphing modes. a, Schematic of the folding behaviours of two-panel devices when actuated using an applied out-of-plane magnetic field. These are achieved by setting the magnetization perpendicular, parallel or at an angle to the folding crease, as indicated by the red arrows on the panels. b, Schematic of the panel design using type I–IV nanomagnets with the magnet long axis oriented along x and y directions and the colour of the nanomagnets corresponding to the colours of the magnetic hysteresis loops in Fig. 1c. c, SEM image of the microscale 'bird'. The coloured bars indicate the location of the arrays of type I-IV nanomagnets and the orientation of the nanomagnets. d, Schematic (left) and optical images (right) of a microscale 'bird' mimicking four flying modes, 'flapping', 'hovering', 'turning' and 'side-slipping'. From left to right in each row: encoded magnetization direction (coloured arrows) of the type I-IV nanomagnets with the colours corresponding to the colours of the magnetic hysteresis loops in Fig. 1c; schematic of a flat microscale 'bird' with the total magnetization direction for each panel indicated with red arrows; schematic showing the

folding of the microscale 'bird' under the indicated controlling field  ${\bf B}$  ; and the optical microscope images of the experimental demonstrations. For 'flapping', the three optical images show the shape transformation in different controlling fields B. For 'hovering', shape transformations in a magnetic field (1.5 mT, 1 Hz), which rotates back and forth, are shown with the three optical images corresponding to three successive field directions during the field rotation. For 'turning', three successive snapshots of the shape transformations in an alternating magnetic field (11.6 mT, 24.5 Hz) are shown, and for 'side-slipping', three successive snapshots of the shape transformation in an alternating magnetic field (6.2 mT, 19.5 Hz) are shown. The solid black arrows indicate the direction of the applied magnetic field. In the optical images of the 'turning' and 'side-slipping' modes, a dashed red line connects a fixed reference point on the substrate (large red dot) and the middle point between the two wings of the 'bird' (small red dot), highlighting the motion of the 'bird'. Scale bars:  ${f c}$ , 15  ${\mu}m$ ; d. 30 um.

was achieved using different combinations of arrays of the four types of nanomagnet, I-IV (Fig. 2b). The micromachine was then constructed from these four assembly units, corresponding to the arrangement of the coding states shown in Fig. 2a. The micromachine is then encoded so that one set of magnetic configurations of the units represents 'P' and the other represents 'X' (Fig. 2c). After releasing the encoded devices from the substrate, the micromachines display shape-morphing into 'P' and 'X' patterns when actuated by an applied controlling field (see Fig. 2d and Supplementary Video 3). In addition to these two patterns, this micromachine design also has '9' and '0' modes, which are conjugate modes of 'P' and 'X' modes, respectively (see Extended Data Fig. 6). This modular design concept can therefore be used to create complex shapemorphing systems with tailored three-dimensional (3D) transformations by customizing the design and layout of the functional units.

More advanced folding behaviours, such as the bending and twisting shown in Fig. 3a, can be achieved by programming the rigid panels of the shape-morphing micromachines with arbitrary magnetic configurations. The magnetic moment of the panels can be tuned by including arrays of nanomagnets with different aspect ratios and orientations (see Methods section 'Design of nanomagnet arrays'). In the example shown in Fig. 3b, with four types of nanomagnet I-IV oriented along both the x and y directions, there are  $2^4 \times 2^4 = 2^8$  magnetic configurations on a single panel. By adjusting the quantities  $n_x$  and  $n_y$ of each type of nanomagnet along the x and y directions, it is possible to assign to each panel a total magnetic moment of almost arbitrary magnitude and direction. As a demonstration, we have engineered an origami-like microscale 'bird' to mimic the different flying modes of a real bird (see Fig. 3c). This is achieved with specific arrangements of the nanomagnet arrays on the five body parts—head, neck, body, tail and a pair of wings. For the tips and joints of the bird's wings, there are four arrays of nanomagnets on each panel with different orientations and switching fields. Therefore the total magnetization on these panels has eight possible directions and one demagnetized state (see Extended Data Fig. 7). By encoding the nanomagnets on the microscale 'bird' with different magnetic configurations, we demonstrate four distinct morphological transformations-'flapping', 'hovering', 'turning' and 'side-slipping' (see Fig. 3d and Supplementary Videos 4-7)—achieved by varying the magnetic fields, as indicated in Fig. 3d. Here advanced folding behaviours are demonstrated; for the 'hovering' mode, the left and right wings 'twist' relative to the body, and in the 'side-slipping' mode, the right wing tip 'bends and twists' relative to the right wing joint. Ways to achieve additional transformations are discussed in Methods section 'Further transformations'.

For future applications, our micromachines have a wide range of tunability in terms of size, ranging from submicrometre-sized panels with a single nanomagnet up to millimetre-sized devices: this range is limited only by the fabrication methods available. The nanomagnet arrays can be further engineered to have temperature-dependent magnetic properties<sup>24</sup>, and they can be modulated using radio-frequency magnetic fields<sup>25</sup> and light<sup>26</sup>. This possibility of control with several different stimuli provides further functionality of the micromachines in many different environments. Nanoscale magnets switch in only a few nanoseconds<sup>25</sup>, which is much faster than the mechanical response of the micromachines that occurs on the millisecond timescale<sup>15</sup>. Therefore, the micromachines can be reprogrammed in situ using a short (nanosecond to millisecond) magnetic field pulse. With the ability to precisely control transformations at the micrometre scale, our micromachines also offer a platform to construct 3D magnetic metamaterials<sup>27</sup>, such as a 3D realization of artificial spin ice<sup>28</sup>, and photonic metamaterials<sup>29</sup>, where optical properties, such as the polarization of transmitted light, can be tuned by magnetically actuated transformations. This concept can also be applied in flexible electronics, with morphable 3D structures having multistable states<sup>30</sup>. By encoding the nanomagnets, the devices can be readily switched between these states using an applied magnetic field.

### **Online content**

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information,

acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1713-2.

- Nelson, B. J., Kaliakatsos, I. K. & Abbott, J. J. Microrobots for minimally invasive medicine. Annu. Rev. Biomed. Eng. 12, 55–85 (2010).
- Li, J., Esteban Fernández De Ávila, B., Gao, W., Zhang, L. & Wang, J. Micro/nanorobots for biomedicine: delivery, surgery, sensing, and detoxification. Sci. Robot. 2, eaam6431 (2017).
- Rus, D. & Tolley, M. T. Design, fabrication and control of soft robots. Nature 521, 467–475 (2015).
- 4. Rich, S. I., Wood, R. J. & Majidi, C. Untethered soft robotics. Nat. Electron. 1, 102-112 (2018).
- 5. Palagi, S. & Fischer, P. Bioinspired microrobots. Nat. Rev. Mater. 3, 113-124 (2018).
- Hu, C., Pané, S. & Nelson, B. J. Soft micro- and nanorobotics. Annu. Rev. Control. Robot. Auton. Syst. 1, 53–75 (2018).
- Zheludev, N. I. & Kivshar, Y. S. From metamaterials to metadevices. Nat. Mater. 11, 917–924 (2012).
- Xia, F. & Jiang, L. Bio-inspired, smart, multiscale interfacial materials. Adv. Mater. 20, 2842–2858 (2008).
- Miskin, M. Z. et al. Graphene-based bimorphs for micron-sized, autonomous origami machines. Proc. Natl Acad. Sci. USA 115, 466–470 (2018).
- Kim, Y., Yuk, H., Zhao, R., Chester, S. A. & Zhao, X. Printing ferromagnetic domains for untethered fast-transforming soft materials. *Nature* 558, 274–279 (2018).
- Sydney Gladman, A., Matsumoto, E. A., Mahadevan, L., Nuzzo, R. G. & Lewis, J. A. Biomimetic 4D printing. Nat. Mater. 15, 413–418 (2016).
- Huang, H.-W., Sakar, M. S., Petruska, A. J., Pane, S. & Nelson, B. J. Soft micromachines with programmable motility and morphology. Nat. Commun. 7, 12263 (2016).
- Magdanz, V., Guix, M., Hebenstreit, F. & Schmidt, O. G. Dynamic polymeric microtubes for the remote-controlled capture, guidance, and release of sperm cells. Adv. Mater. 28, 4084–4089 (2016)
- Jin, D. et al. Four-dimensional direct laser writing of reconfigurable compound micromachines. Mater. Today https://doi.org/10.1016/j.mattod.2019.06.002 (2019)
- Peyer, K. E., Zhang, L. & Nelson, B. J. Bio-inspired magnetic swimming microrobots for biomedical applications. *Nanoscale* 5, 1259–1272 (2013).
- Erb, R. M., Martin, J. J., Soheilian, R., Pan, C. & Barber, J. R. Actuating soft matter with magnetic torque. Adv. Funct. Mater. 26, 3859–3880 (2016).
- Hu, W., Lum, G. Z., Mastrangeli, M. & Sitti, M. Small-scale soft-bodied robot with multimodal locomotion. *Nature* 554, 81–85 (2018).
- Xu, T., Zhang, J., Salehizadeh, M., Onaizah, O. & Diller, E. Millimeter-scale flexible robots with programmable three-dimensional magnetization and motions. Sci. Robot. 4, eaav4494 (2019).
- Kim, J. et al. Programming magnetic anisotropy in polymeric microactuators. Nat. Mater. 10, 747–752 (2011).
- Huang, T. Y. et al. 3D Printed microtransporters: compound micromachines for spatiotemporally controlled delivery of therapeutic agents. Adv. Mater. 27, 6644–6650 (2015).
- Cowburn, R. P., Koltsov, D. K., Adeyeye, A. O. & Welland, M. E. Single-domain circular nanomagnets. Phys. Rev. Lett. 83, 1042–1045 (1999).
- Cowburn, R. P. Property variation with shape in magnetic nanoelements. J. Phys. D 33, R1–R16 (2000).
- Rus, D. & Tolley, M. T. Design, fabrication and control of origami robots. Nat. Rev. Mater. 3, 101–112 (2018).
- Gliga, S. et al. Emergent dynamic chirality in a thermally driven artificial spin ratchet. Nat. Mater. 16, 1106–1111 (2017).
- Barros, N., Rassam, M., Jirari, H. & Kachkachi, H. Optimal switching of a nanomagnet assisted by microwaves. *Phys. Rev. B* 83, 144418 (2011).
- Lambert, C. H. et al. All-optical control of ferromagnetic thin films and nanostructures Science 345, 1337–1340 (2014).
- Fernández-Pacheco, A. et al. Three-dimensional nanomagnetism. Nat. Commun. 8, 15756 (2017).
- Heyderman, L. J. & Stamps, R. L. Artificial ferroic systems: novel functionality from structure, interactions and dynamics. J. Phys. Condens. Matter 25, 363201 (2013).
- Gansel, J. K. et al. Gold helix photonic metamaterial as broadband circular polarizer. Science 325, 1513–1515 (2009).
- Fu, H. et al. Morphable 3D mesostructures and microelectronic devices by multistable buckling mechanics. Nat. Mater. 17, 268–276 (2018).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

### **Methods**

### Sample fabrication

A schematic of the fabrication process can be found in Extended Data Fig. 8. The samples were fabricated on a 50-nm-thick low-stress silicon nitride membrane (Silson Ltd, UK) supported by a rigid silicon frame. First, the nanomagnets were fabricated on the membrane with electron beam lithography using an electron beam writer (Vistec EBPG 5000PlusES) to pattern a spin-coated 50k poly(methylmethacrylate) (PMMA)/950k PMMA double-layer. A magnetic film of 5 nm Ti (adhesion layer)/60 nm Co/3 nm Al (capping layer) was thermally evaporated at a base pressure of  $\sim 1 \times 10^{-6}$  mbar onto the patterned resist, which was followed by a lift-off process in acetone. Then the panels were fabricated by spin-coating a 950k PMMA layer on the front of the silicon nitride membrane, which was then patterned using electron beam lithography in order to define the geometry of the rigid panels and hinge springs of the micromachines. After coating a second 950k PMMA layer on the back of the membrane, reactive ion etching (RIE, Oxford PlasmaPro 100) was performed on the front side to etch through the 50-nm-thick silicon nitride membrane. After the RIE process, the fabricated micromachines were only supported by the free-standing PMMA layer coated on the back of the membrane. The devices were then magnetized by a sequence of magnetic fields and released in an organic solvent, propylene glycol monomethyl ether acetate (PGMEA). A further discussion of this release process and possible operation of the micromachines in other media can be found in Methods section 'Micromachine release and operation'. The SEM (Zeiss Supra VP55) images were taken with a 3-10 kV acceleration voltage.

### Magnetic characterization and encoding

The nanomagnets were characterized with MOKE measurements using a commercial setup (NanoMOKE, Durham Magneto Optics Ltd.) in the longitudinal mode with a focused laser spot with a diameter of ~10  $\mu m$ . Each hysteresis loop was obtained by averaging ten measurements. The nanomagnet arrays were encoded using the magnetizing field from the electromagnet in the NanoMOKE setup. The electromagnet was equipped with a manual rotation stage to hold the sample between the electromagnet iron poles. By rotating the sample on this stage, a magnetic field of up to 400 mT could be applied in any direction in the sample plane.

### Magnetic manipulation

After the sample fabrication and the magnetic field encoding, the sample was released in an organic solvent (PGMEA) while being observed with an optical microscope incorporating three pairs of Helmholtz coils in an orthogonal configuration. A computer was used to simultaneously control the electric currents in the three-pair Helmholtz coils to produce a magnetic field vector in arbitrary 3D directions. A magnetic field with a maximum magnitude of 15 mT was generated to manipulate the released micromachines. The observed folding behaviour and motion were captured with a video camera (Grasshopper GRAS-03K2C, Point Grey Research) on the microscope. The tested micromachines were actuated reversibly with a dynamic field more than 18,000 times (30 Hz, 10 min) with no observable signs of plastic deformation or breaking.

### Nanomagnet design and coercivity

The magnetic domain state of an element made of a soft polycrystal-line ferromagnetic material (typically Fe, Ni and Co) is dictated by the competition between the quantum mechanical exchange energy and the stray field energy that is size- and shape-dependent<sup>31</sup>. By varying the lateral size and shape, it is possible to create single-domain nanomagnets.

The nanomagnets are designed in a stadium shape, comprising a rectangle with one semi-circle at each end, with the width of the rectangle matching the semi-circle diameter, as shown in Extended Data

Fig. 2a. This shape ensures that the remanent magnetization is parallel to the long axis of the magnets, pointing in one of the two directions. The volume of the stadium-shaped nanomagnet is given by:

$$V = \left[ d(L - d) + \pi \left( \frac{d}{2} \right)^2 \right] t \tag{1}$$

When the length, width and thickness of the nanomagnets are L=300 nm, d=110 nm and t=60 nm, respectively, the volume of the nanomagnet is  $V_0=1.82\times 10^{-21}$  m³. The total magnetic moment of a nanomagnet is given by m=MV, where M is the saturation magnetization of the magnetic material. In this work, 60-nm-thick cobalt nanomagnets are employed, and the saturation magnetization of the thermally evaporated Co thin film is  $M_s=1,153$  kA m⁻¹, measured using a superconducting quantum interference device vibrating sample magnetometer (SQUID VSM) at room temperature.

Nanomagnets with different aspect ratios, L/d, display different magnetic coercivities, as shown in Fig. 1c. Here we keep the same nanomagnet volume  $V_0$  and thickness  $t_0$  = 60 nm for all nanomagnets regardless of the aspect ratio. Therefore, they have the same magnetic moment  $m_0$  =  $MV_0$  and equation (1) can be modified to read:

$$V_0 = \left[ d(L - d) + \pi \left( \frac{d}{2} \right)^2 \right] t_0$$
 (2)

So that the relation between L and d is given by:

$$L = d + \frac{\frac{V_0}{t_0} - \pi \left(\frac{d}{2}\right)^2}{d} \tag{3}$$

A plot of d against L, determined using equation (3), is shown in Extended Data Fig. 2c. In this figure, nanomagnets with the same volume  $V_0$  but with six different aspect ratios are selected, that is, with lateral dimensions of  $520 \, \mathrm{nm} \times 60 \, \mathrm{nm}$ ,  $450 \, \mathrm{nm} \times 70 \, \mathrm{nm}$ ,  $398 \, \mathrm{nm} \times 80 \, \mathrm{nm}$ ,  $358 \, \mathrm{nm} \times 90 \, \mathrm{nm}$ ,  $326 \, \mathrm{nm} \times 100 \, \mathrm{nm}$  and  $300 \, \mathrm{nm} \times 110 \, \mathrm{nm}$ . The layout of the nanomagnet arrays is schematically shown in Extended Data Fig. 2b, with a spacing of d/2 in the x direction. In this layout, the dipolar coupling between the nanomagnets supports parallel alignment of the magnetization between neighbouring magnets. In the y direction, neighbouring nanomagnets are separated by a distance  $s = 40 \, \mathrm{nm}$  for all nanomagnet arrays.

SEM images of the fabricated arrays are given in Extended Data Fig. 2e. The MOKE magnetic characterization is shown in Fig. 1c, with the square shapes of the hysteresis loops confirming the single-domain magnetic state of the nanomagnets. Since all nanomagnets investigated in this study have same magnetic moment,  $m_0 = 2.10 \times 10^{-15} \, \mathrm{A} \, \mathrm{m}^2$ , the total magnetic moment of the nanomagnet arrays on an individual panel (when all magnetizations are aligned in the same direction) is

$$m_{\text{total}} = n_{\text{magnets}} m_0$$
 (4)

where  $n_{\rm magnets}$  is the number of nanomagnets on the panel. For the device demonstrated in Fig. 1, a total of 1,040 nanomagnets are fabricated on each panel, with  $m_{\rm total} = 2.18 \times 10^{-12} \, {\rm A m^2}$  pointing parallel to the long axis of the nanomagnets.

The square shapes of the hysteresis loops measured using the MOKE, shown in Fig. 1c, indicate that all six types of designed nanomagnets are fully magnetized at remanence. As shown by the MOKE curves in Extended Data Fig. 2d, the magnetic switching occurs over a relatively small field range of 5–20 mT. Since the six transition regions do not overlap, all six types of nanomagnets can be individually programmed, even for a micromachine containing arrays of all six nanomagnets oriented in the same direction.

### Hinge spring design and properties

Inspired by the art of origami, we have designed rigid panels carrying single-domain nanomagnets, connected by structured hinge springs acting as folding creases. The hinge spring layout needs to be designed so that there is a considerable folding behaviour on the application of a small magnetic field, B < 15 mT.

We first derive the relationship between applied magnetic field B and panel rotation angle  $\theta$ . Neglecting the bending of the short sections of the spring, the twisting of the zig-zag spring can be determined by considering a long and slender beam of equivalent length that can twist subject to an applied torque.

Considering first the twist of an isolated beam with uniform cross-section along its length (see Extended Data Fig. 3a), the angle of twist in radians is given by  $\theta = \frac{\tau L}{GJ}$ , where  $\tau$  is the applied torque, L is the beam length, and G and J are the shear modulus and torsional constant of the material, respectively.

For a beam with a rectangular section, the torsional constant is given by

$$J = \frac{wt(w^2 + t^2)}{12} \tag{5}$$

where w and t are the side lengths<sup>32</sup>. For a homogeneous isotropic material,

$$G = \frac{E}{2(1+\nu)} \tag{6}$$

where E is the Young's modulus and v is the Poisson ratio. Hence the panel rotation angle is given by:

$$\theta = \frac{24(1+v)L}{Ewt(w^2+t^2)}\tau$$
 (7)

The applied torque is given by  $\tau = k\theta$ , so that the magnitude of the torsional spring constant, k, is given by

$$k = \frac{Ewt(w^2 + t^2)}{24(1+v)L} \tag{8}$$

For the current design, the total length of such a beam is L = nl, where l is the length of each beam section and n is the total number of beam sections. Therefore, for each spring:

$$k_{\rm s} = \frac{Ewt(w^2 + t^2)}{24(1+v)nl} \tag{9}$$

If two springs are used and they are in a parallel configuration:

$$k_{\text{total}} = k_{\text{s}} + k_{\text{s}} = \frac{Ewt(w^2 + t^2)}{12(1 + v)nl}$$
 (10)

Taking into account that the direction of panel rotation  $\theta$  is opposite to the direction of the mechanical torque induced on the spring, the total torque on the spring is given by:

$$\tau_{\text{spring}} = -k_{\text{total}}\theta \tag{11}$$

We now determine the extent of the folding in a magnetic field for hinge designs with different numbers of turns in our micromachines. For this, we have fabricated simple two-panel devices to investigate the relationship between the applied magnetic field and the panel rotation angle. An SEM image of an eight-turn two-panel device is shown in Extended Data Fig. 3b. The fabricated devices have identical arrays of

nanomagnets (398 nm  $\times$  80 nm nanomagnets on the left panel and 520 nm  $\times$  60 nm nanomagnets on the right panel) but the number of turns in the hinge spring is varied, with q=2,4,6 and 8; see Extended Data Fig. 3c. The total length of each spring is therefore L=nl=(2q+1)l, where l=5  $\mu$ m is the length of each section of the spring.

According to the illustration in Extended Data Fig. 3d, the magnetic-field-induced torque on the device is  $\tau_B = m \times B$  and its magnitude is:

$$\tau_B = mB\sin(90^\circ - \theta) = mB\cos\theta \tag{12}$$

At equilibrium:

$$\tau_B + \tau_{\text{spring}} = 0 \tag{13}$$

Introducing equations (11) and (12) into equation (13) gives:

$$mB \cos(\theta) - k_{\text{total}}\theta = 0$$

$$\frac{\theta}{\cos\theta} = \frac{m}{k_{\text{total}}} B \tag{14}$$

Introducing equations (4) and (10) into equation (14), we obtain:

$$\frac{\theta}{\cos\theta} = \frac{n_{\text{magnets}} m_0 12(1+v)nl}{Ewt(w^2 + t^2)} B$$
 (15)

This equation is the relation between applied magnetic field H and panel rotation angle  $\theta$ , which can be numerically solved. Here  $n_{\text{magnets}} = 1,175, n = [5, 9, 13, 17]$  for devices with 2-, 4-, 6- and 8-turn hinges. and the lateral dimensions are w = 100 nm, t = 50 nm and  $l = 5 \mu m$ , as  $used in the \, experimental \, micromachines. \, From \, equation \, (4), the \, total \,$ magnetic moment on each panel is  $m_{\text{total}} = 2.47 \times 10^{-12} \,\text{A m}^2$ . Shown in Extended Data Fig. 3e are optical microscope images of the panel rotation for hinged spring designs with different numbers of turns following actuation in the same field,  $B = 5 \,\mathrm{mT}$ . The experimental demonstrations and computational results are plotted in Extended Data Fig. 3f, g, respectively, and follow similar trends. The panel rotation angle is larger in the experiment, which may be partly due to overetching in the final nanofabrication step leading to hinged springs that are narrower than in the original design. For the devices shown in Figs. 1–3, hinge springs with a = 6 turns were used. With the current possibilities for magnetic field control of field steps of 0.1 mT, we can orient the panels to within ~0.1° to 1°, depending on the panel rotation angle in an applied magnetic field.

We now estimate the magnitude of the largest torque and force that can be generated from the panels patterned with nanomagnets. The induced mechanical torque on the panels  $\tau$  has a linear relationship with the magnitude of the applied magnetic field B, since  $\tau = \mathbf{m} \times \mathbf{B}$ . However, on increasing B, the mechanical torque  $\tau$  cannot be infinitely high, since the applied magnetic field may switch the nanomagnets. For example, if the magnetic field is applied along the long axis of the magnets, then it can alter the magnetization if it is larger than the switching field, which is 30 mT for the type IV nanomagnets.

For a given micromachine, there are several panels in different orientations that change as the magnetic field is applied. We therefore assume that we can safely operate a machine with an upper limit to the field given by the coercive field  $B_c$ , that is, the switching field of the magnets when applying a field parallel to the magnet long axis.

We calculate the torque and force that can be generated by a panel patterned with type IV nanomagnets. For the four-panel devices and their modular assemblies in Figs. 1, 2, there are in total 1,040 nanomagnets on each 10  $\mu$ m × 10  $\mu$ m panel, with  $m_{\text{total}} = 2.18 \times 10^{-12} \, \text{A} \, \text{m}^2$  pointing along the long axis of the nanomagnets. In this case, for a 30 mT operation field, the highest generated magnetic torque is  $\tau = (2.18 \times 10^{-12} \, \text{A} \, \text{m}^2) \times (30 \, \text{mT}) \times \sin(90^\circ) = 6.54 \times 10^{-14} \, \text{N} \, \text{m}$ . Assuming one edge of the panel

is hinged, then the highest generated force is  $f = (6.54 \times 10^{-14} \text{ N m})/(10 \text{ µm}) = 6.54 \times 10^{-9} \text{ N} = 6.54 \text{ nN}.$ 

Similarly, for a  $10 \, \mu m \times 10 \, \mu m$  panel patterned with type I nanomagnets, which have a switching field of 140 mT, the highest generated mechanical torque is  $\tau = 3.05 \times 10^{-13} \, N$  m, and the highest generated force is  $f = 30.5 \, nN$ .

It has been shown previously that microrobots can be used to move cells by pushing them with forces in the piconewton range  $^{33}$ . For moving even larger objects, such as a  $100~\mu m \times 10~\mu m \times 8~\mu m$  microbar, it has been shown that -200 pN of force is needed  $^{34}$ . Therefore, since our machines can supply forces up to several nanonewtons, they are capable of manipulating biological objects such as cells.

#### Design of nanomagnet arrays

We engineer the orientation and magnitude of the magnetic moment of the arrays of nanomagnets. Here, we can consider a panel consisting of nanomagnets with p different aspect ratios, each with different switching fields  $B_{c,i}$  and magnetic moments  $m_i = M_i V_i$ , where i = 1, 2, ..., p.  $M_i$  is the magnetization of the magnetic material and  $V_i$  is the volume of the nanomagnet. The nanomagnets with magnetic moment  $m_i$  can be fabricated with several different arbitrary orientations  $\phi_{i,j}$  on a panel. Here,  $j = 1, 2, ..., k_i$ , where  $k_i$  denotes the total number of different orientations for the nanomagnets with magnetic moment  $m_i$ . Therefore, we obtain the total magnetic moment  $m_{\text{total}}$  for all arrays of nanomagnets on a given panel

$$m_{\text{total}} = \sum_{i=1}^{p} \sum_{j=1}^{k_i} n_{i,j} m_i (\cos \varphi_{i,j} + i \sin \varphi_{i,j})$$
 (16)

where  $n_{i,j}$  is the quantity of the nanomagnets with magnetic moment  $m_i$  and orientation  $\phi_{i,j}$ . This total magnetic moment,  $m_{\text{total}}$ , of nanomagnets on a single panel can be programmed with arbitrary magnitude and direction by careful adjustment of the magnetic moment  $m_i = M_i V_i$ , orientations  $\phi_{i,j}$  and quantities  $n_{i,j}$  of the nanomagnets.

In our micromachines, both the magnetization  $M_i$  and the volume  $V_i$  of the nanomagnets are the same for nanomagnets with different switching fields, so we can assume that magnitude of the magnetic moment  $m_i = M_i V_i$  of all magnets in the array are the same, that is,  $m_i = M_i V_i = M_0 V_0 = m_0$ . Therefore, equation (16) can be simplified to:

$$m_{\text{total}} = m_0 \sum_{i=1}^{p} \sum_{j=1}^{k_i} n_{i,j} (\cos \varphi_{i,j} + i \sin \varphi_{i,j})$$
 (17)

In our systems, the nanomagnets only have two orientations, which are orthogonal to each other. This means that the magnets in the two orthogonal arrays can be magnetized independently by applying a field parallel to the long axes of the nanomagnets in one of the arrays. To be specific, in our design, the long axis of each magnet is aligned along one of two orientations, that is, parallel to one of the two coordinate axes (x or y axis, so  $k_i = 2$ ). Therefore, the magnetization of the nanomagnets can point along one of four cardinal directions (North, South, East and West) so that the orientation of the total magnetic moment is given by

$$\varphi = \tan^{-1} \frac{\sin \varphi_{i,y} \sum_{i=1}^{p} n_{i,y}}{\cos \varphi_{i,y} \sum_{i=1}^{p} n_{i,x}}$$
(18)

where  $n_{i,x}$ ,  $n_{i,y}$  are the number of i types of nanomagnets with a given switching field with the long axis along the x and y axis, respectively, so that  $\varphi_{i,x}=0$ ,  $\pi$  and  $\varphi_{i,y}=\pm\frac{\pi}{2}$  depending on the direction of magnetization in the nanomagnets. According to equation (18), by changing the number of magnets with a given orientation, that is,  $n_{i,x}$  and  $n_{i,y}$ , it is possible to obtain an arbitrary orientation  $\phi$  of the total magnetic moment  $m_{\text{total}}$ .

For each panel, there are nanomagnets with  $p_x$  and  $p_y$  different switching fields along the x and y axis, respectively. It is therefore possible to program  $2^{p_x} \times 2^{p_y}$  magnetic configurations, which have a total magnetic moment with a particular magnitude and orientation.

In order to encode a panel with two or more different types of nanomagnet, one starts by magnetizing the highest-coercivity magnet, so that all magnets on the panel will be magnetized in the direction of the field. If required, the lower-coercivity magnets can be oppositely magnetized using a magnetic field applied in the reverse direction that is sufficiently low to leave the higher-coercivity nanomagnets unaffected.

#### **Further transformations**

In this article, the transformations are inspired by origami, and the micromachines are constructed with rigid panels and structured soft creases. The rigid panels carry the programmed nanomagnet arrays, which provide the mechanical torque for transformations when an external magnetic field is applied. Here, the nanomagnets are designed to be stadium-shaped, with the remanent magnetization pointing in the plane of the rigid panel along the geometric long axis of the magnets as a result of the magnetic shape anisotropy.

The current concept has limitations where folding into particular shapes is not straightforward. For example, the four-panel devices shown in Fig. 1, which have all four side panels magnetized towards or away from the centre, can achieve folding into an 'uncovered box'. Shutting the lid to give a 'closed' box with an additional panel is not possible with the arrangement of nanomagnets used in this work. Instead, the magnetization of the nanomagnets on the lid needs to point out of the plane of the lid. This requires nanomagnets with out-of-plane magnetic anisotropy, which can be achieved with, for example, a Co/Pt multilayer system<sup>35</sup>.

#### Micromachine release and operation

We used organic solvents, such as PGMEA and acetone, to dissolve the PMMA supporting layer and release the micromachines from the silicon substrate in the final step of the fabrication process. We chose PGMEA because it does not evaporate as fast as acetone and, in order to keep the experimental processes simple, we directly actuated the micromachines in the PGMEA solvent after the release. Nevertheless, the micromachines can also be operated in other working environments such as water and air and, since the nanomagnets have a 3 nm Al capping layer, they will not be easily oxidized. This 3 nm Al layer also ensures the biocompatibility of the micromachines. Here, we suggest three approaches to transfer and operate the structures in a water environment:

**Approach 1.** Instead of using a PMMA support layer coated on the back of the membrane (steps 3–4 in Extended Data Fig. 8), water-soluble coatings, such as poly(acrylic acid) or Dextran<sup>36</sup>, can be used. These are compatible with microfabrication techniques and, after the RIE etching, the structures can be directly released and operated in water.

**Approach 2.** After the fabrication and the release of the structures in an organic solvent, the structures can be fished out and transferred to a water environment using a micromanipulator tip. This is a standard procedure for micro- and nano-robotic manipulation, and has been widely reported<sup>20</sup>.

Approach 3. After the fabrication and the release of the structures in an organic solvent, water substitution can be performed, for example, in a Petri dish containing the solvent and the micromachines. This technique is widely used in soft or shape-morphing microrobotics<sup>14</sup>. It should be noted that PGMEA has limited miscibility with water. Therefore, if PGMEA is used to dissolve the PMMA for device release, acetone or isopropyl alcohol (IPA; fully miscible with PGMEA) can be used to substitute PGMEA first, and then a water substitution can be performed to replace acetone or IPA. An alternative approach is to directly use acetone to dissolve PMMA for device release, followed by a water substitution.

Using approach 3 with acetone to dissolve the PMMA, we have demonstrated operation of the four-panel device in water (see Extended Data Fig. 9a, b). In addition, we have demonstrated the manipulation of 6- $\mu$ m-diameter polystyrene microbeads (Polybead 15714-5, Polysciences, Inc.) in a water environment (Extended Data Fig. 9c, d and Supplementary Video 8).

We have also demonstrated the operation of the micromachines in air. For this, we have fabricated single panels with one end attached by a hinge spring connection to a fixed silicon nitride membrane frame. In this case, both the PMMA support layer and the device release in PGMEA are not required, and the device is free to be manipulated in a magnetic field immediately after the reactive ion etching process. As shown in Extended Data Fig. 10 and Supplementary Video 9, on increasing an applied out-of-plane magnetic field, the panels fold upwards or downwards depending on the programmed magnetization direction of the nanomagnet arrays on each panel.

It is worth noting that, even without suspension in a solvent, the micromachines will not collapse due to the gravitational force. For the four-panel micromachine shown in Fig. 1, the weight is the combined weight of the silicon nitride membrane and the nanomagnets. The weight of silicon nitride panels is approximately given by  $\rho_{SiNx} \times$  $V_{\text{SiNx}} \times g = 7.8 \times 10^{-13} \,\text{N}$ , where  $\rho_{\text{SiNx}} = 3.17 \,\text{g cm}^{-3}$  is the density of silicon nitride,  $V_{\text{SiNx}} = 5 \times 10 \, \mu\text{m} \times 10 \, \mu\text{m} \times 50 \, \text{nm}$  is the volume of the fourpanel micromachine consisting of 5 rigid panels, and  $g \approx 9.8 \text{ m s}^{-2}$  is the standard gravity. The weight of the nanomagnets is approximately  $\rho_{\rm magnets}$  ×  $V_{\rm magnets}$  × g = 6.6 ×  $10^{-13}$  N, where  $\rho_{\rm magnets}$  = 8.9 g cm $^{-3}$  is the cobalt density,  $V_{\text{magnets}} = 4 \times 1,040 \times V_0$  is the total volume of the nanomagnets on the micromachine with four side panels, each having 1,040 nanomagnets, and  $V_0 = 1.82 \times 10^{-21} \,\mathrm{m}^3$  is the volume of each individual nanomagnet. Therefore the total weight of the four-panel micromachine is  $7.8 \times 10^{-13} \,\text{N} + 6.6 \times 10^{-13} \,\text{N} = 1.44 \,\text{pN}$ . The magnitudes of the magnetic actuation force and the elastic force from the hinge springs are in the nanonewton range (as calculated in the Methods section 'Hinge spring design and properties'), which is three orders of magnitude larger than the gravitational load. Therefore the devices can be manipulated in air without structural collapse due to gravity.

#### **Data availability**

All data generated or analysed during this study are included in the published article and its Supplementary Information, and are available from the corresponding authors on reasonable request.

- 31. O'Handley, R. C. Modern Magnetic Materials: Principles and Applications (Wiley, 1999).
- 32. Timošenko, S. P. & Goodier, J. N. Theory of Elasticity (McGraw-Hill, 1961).
- Steager, E. B. et al. Automated biomanipulation of single cells using magnetic microrobots. *Int. J. Robot. Res.* 32, 346–359 (2013).
- Huang, T. Y. et al. Cooperative manipulation and transport of microobjects using multiple helical microcarriers. RSC Advances 4, 26771–26776 (2014).
- 35. Luo, Z. et al. Chirally coupled nanomagnets. Science 363, 1435-1439 (2019).
- Linder, V., Gates, B. D., Ryan, D., Parviz, B. A. & Whitesides, G. M. Water-soluble sacrificial layers for surface micromachining. Small 1, 730–736 (2005).

Acknowledgements J.C. received support from the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska Curie Grant Agreement (number 701647). This work was financially supported by the European Research Council Advanced Grant–Soft MicroRobots (SOMBOT, number 743217), the Swiss National Science Foundation (number 200021\_165564) and the National Natural Science Foundation of China (number 11702003). We thank A. Weber and V. Guzenko for helping with the development of the fabrication process. The sample fabrication was performed using the cleanroom facilities at the Laboratory for Micro- and Nanotechnology at the Paul Scherrer Institute, Switzerland.

Author contributions J.C., T.-Y.H., L.J.H. and B.J.N. conceived the project. J.C., Z.L., P.T. and X.-Z.C. developed the fabrication process. T.-Y.H., J.C., z.L. and H.G. developed the design strategy of nanomagnetic encoding. T.-Y.H., J.C. and X.-Z.C. tested the shape-morphing performance. J.C., T.-Y.H. and L.J.H. worked on the manuscript together. All authors contributed to the discussion of the results and the manuscript revision.

Competing interests The authors declare no competing interests.

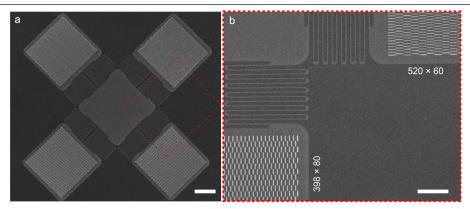
#### Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1713-2

Correspondence and requests for materials should be addressed to T.-Y.H. or J.C.

Peer review information Nature thanks Je-Sung Koh, Xuanhe Zhao, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

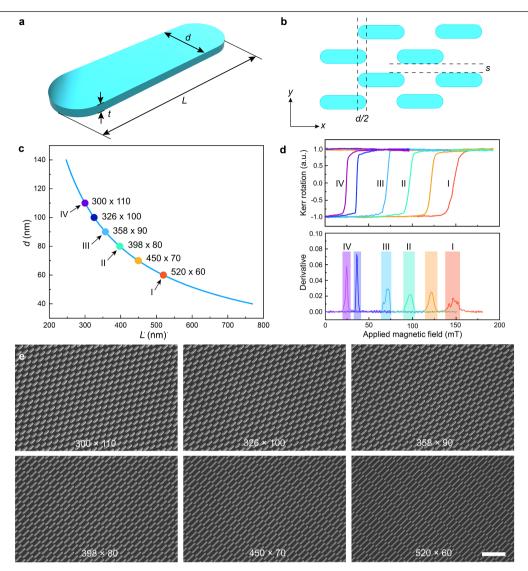
Reprints and permissions information is available at http://www.nature.com/reprints



### Extended Data Fig. 1 | SEM images of a four-panel micromachine.

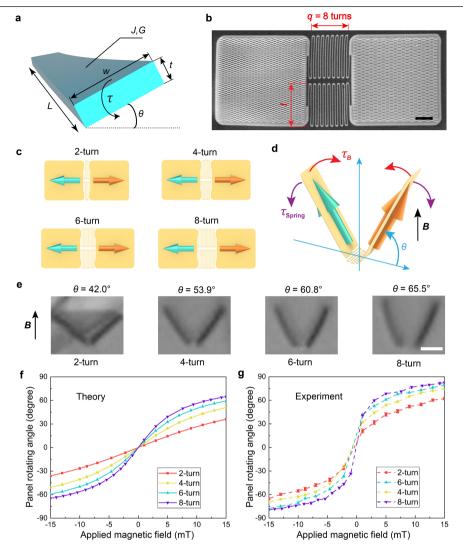
 ${\bf a}, Overview. \, {\bf b}, Enlarged \, image \, corresponding \, to \, the \, dashed \, box \, in \, {\bf a}. \, Shown \, are \, arrays \, of \, nanomagnets: \, in \, the \, array \, at \, top \, right, \, the \, lateral \, dimension \, of \, array \, at \, top \, right, \, the \, lateral \, dimension \, of \, array \, at \, top \, right, \, the \, lateral \, dimension \, of \, array \, at \, top \, right, \, the \, lateral \, dimension \, of \, array \, at \, top \, right, \, the \, lateral \, dimension \, of \, array \, at \, top \, right, \, the \, lateral \, dimension \, of \, array \, at \, top \, right, \, the \, lateral \, dimension \, of \, array \, at \, top \, right, \, the \, lateral \, dimension \, of \, array \, at \, top \, right, \, the \, lateral \, dimension \, of \, array \, at \, top \, right, \, the \, lateral \, dimension \, of \, array \, at \, top \, right, \, the \, lateral \, dimension \, of \, array \, at \, top \, right, \, the \, lateral \, dimension \, of \, array \, at \, top \, right, \, the \, lateral \, dimension \, of \, array \, at \, top \, right, \, the \, lateral \, dimension \, array \, at \, top \, right, \, the \, lateral \, dimension \, at \, array \, at \, top \, right, \, the \, lateral \, dimension \, at \, array \, at \, top \, right, \, the \, lateral \, dimension \, at \, array \, at \,$ 

each nanomagnet is 520 nm  $\times$  60 nm; at bottom left, the lateral dimension of each nanomagnet is 398 nm  $\times$  80 nm. Scale bars: a, 4  $\mu$ m; b, 2  $\mu$ m.



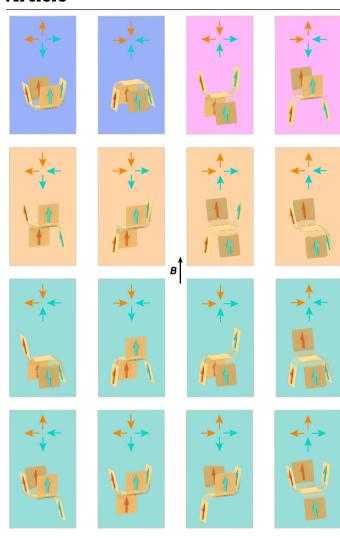
**Extended Data Fig. 2** | **Geometric design and switching behaviour of the nanomagnets. a**, Schematic of a stadium-shaped nanomagnet with length L, width d and thickness t. **b**, Schematic of the layout of the nanomagnet arrays with vertical separation s, and horizontal separation d/2. **c**, Relationship between d and d for nanomagnets with the same volume d0 and thickness d0 nm. Six nanomagnets with different aspect ratios are indicated on the curve (the dimensions of each magnet are indicated in nm); the colour of the points corresponds to colour of the hysteresis loops in Fig. 1c. Arrows indicate

the four types of nanomagnet used in the micromachines (I–IV).  $\mathbf{d}$ , Magneto-optical Kerr effect curves for the six differently sized nanomagnets in the field region where they switch (top panel) and the derivative with the switching region highlighted with shaded boxes (bottom panel). As the six switching regions do not overlap, all six nanomagnets can be individually programmed.  $\mathbf{e}$ , SEM images of fabricated arrays of nanomagnets with lateral dimensions given in nanometres, corresponding to the six coloured points in  $\mathbf{c}$ . Scale bar at bottom right (1  $\mu$ m) applies to all six images.

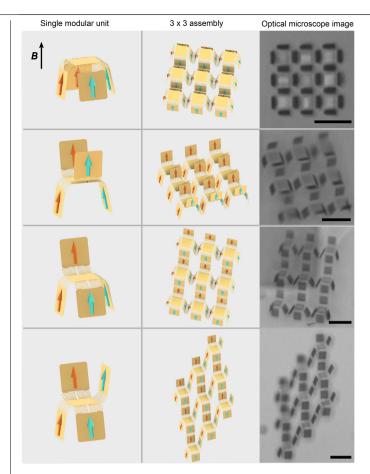


**Extended Data Fig. 3** | **Hinge spring design. a**, Schematic of a single section of a spring. See Methods for nomenclature. **b**, SEM image of a two-panel device with an 8-turn spring. **c**, Schematic of two-panel devices with 2-, 4-, 6- and 8-turn spring designs. The turquoise and orange arrows represent the magnetization direction of the panels. **d**, Schematic of a two-panel device that folds when applying a controlling magnetic field **B**. See Methods for nomenclature. **e**, Optical microscope images of four fabricated devices with different numbers of turns in the spring design on application of a 5 mT

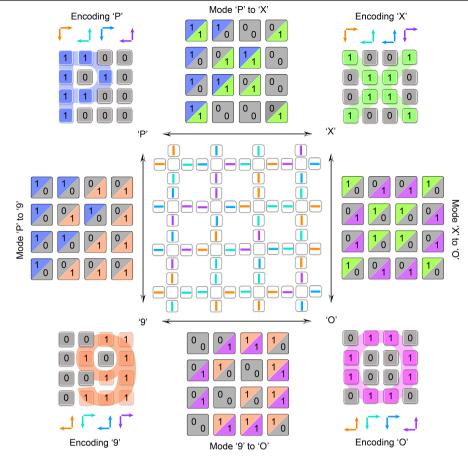
controlling field. **f**, Predicted panel rotating angle versus applied magnetic field based on theoretical calculations of the two-panel devices with different numbers of turns. **g**, Measured panel rotating angle versus applied magnetic field for the fabricated devices with different numbers of hinge spring turns. Each data point corresponds to the average of three measurements of the angle using image analysis software. Error bars,  $\pm 1$  s.d. Scale bars: **b**,  $2\,\mu m$ ; **e** (applies to all images in **e**),  $5\,\mu m$ .



Extended Data Fig. 4 | The 16 magnetization configurations of a four-panel micromachine and their corresponding shape transformation after applying a vertical controlling field. The four background colours highlight the family of four distinct conformations demonstrated in Fig. 1c.

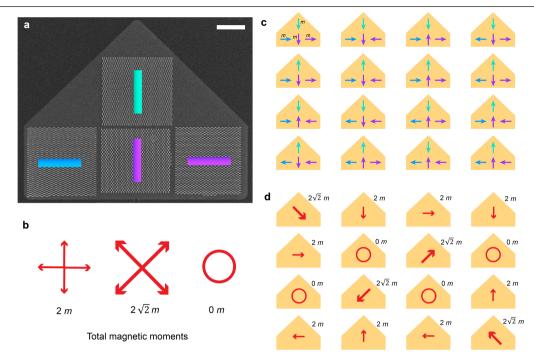


Extended Data Fig. 5 | Four conformations of a micromachine consisting of a  $3\times3$  assembly of four-panel modular units. Shown in the middle and right panels are schematics and experimental demonstrations of the actuated micromachines. The units in a given micromachine all have the same conformation (left panel) corresponding to one of the 16 different magnetization configurations. Scale bars in the optical microscope images,  $30\,\mu m$ .



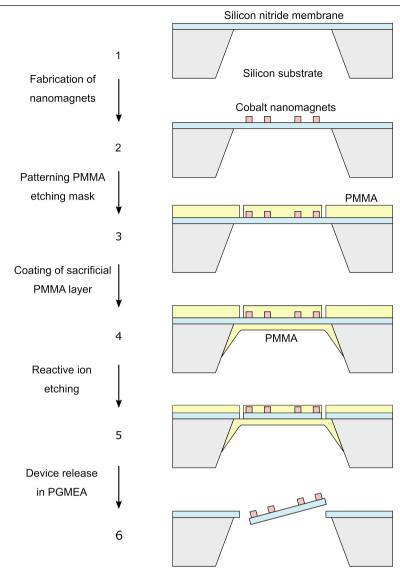
Extended Data Fig. 6 | Four different modes, 'P', 'X', 'O' and '9', encoded in the same micromachine design. In the conjugate pairs ('P' and '9', or 'X' and 'O'), the 'up' and 'down' states of each unit are reversed. With a different

 $nanomagnet\,encoding, a single\,micromachine\,with\,this\,design\,can\,transform\,between\,these\,four\,modes.\,See\,main\,text\,and\,Methods\,for\,details.$ 



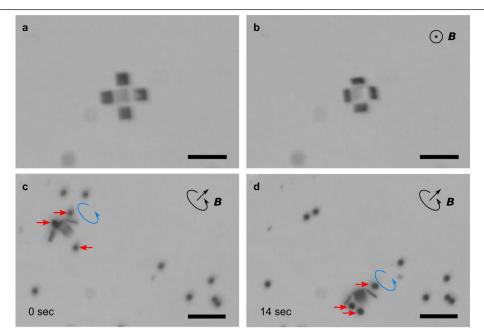
Extended Data Fig. 7 | Possibilities for the total magnetic moments of the wing tip in the microscale 'bird'. a, SEM image of the wing tip of the microscale bird. Turquoise vertical bar, type II nanomagnets (398 nm  $\times$  80 nm); blue horizontal bar, type III nanomagnets (358 nm  $\times$  90 nm); purple bar (horizontal and vertical), type IV nanomagnets (300 nm  $\times$  110 nm). Each of the arrays has the same number of magnets (1,040) with the same magnetic moment m. b, Nine possible total magnetic moment magnitudes and directions. c, Schematics of 16 possible magnetic configurations of the wing tip. Each of the arrays of different types of nanomagnets (types II, III and IV) have different

switching fields, and there are two out of the four arrays that have the same type IV magnets but with orthogonal orientation. Therefore, with the orientation of the nanomagnets in two of the arrays along the x direction and in the two other arrays along the y direction, there are in total  $2^2\times 2^2=16$  possible magnetic configurations that can be encoded into the wing tip.  $\boldsymbol{d}$ , Schematics showing the magnitudes and directions of the total magnetic moment of the wing tip, corresponding to the 16 magnetic configurations shown in  $\boldsymbol{c}$ . Scale bar in  $\boldsymbol{a}$ ,  $4\,\mu\text{m}$ .



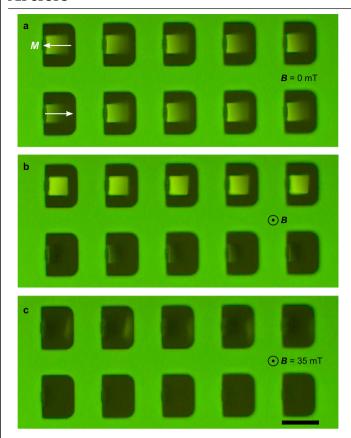
 $\label{lem:extended} \textbf{Extended Data Fig. 8} | \textbf{Schematic of the steps used to fabricate the micromachines.} The nanomagnets are fabricated using electron beam lithography, including patterning of a spin-coated polymer resist, thermal lithography including patterning of a spin-coated polymer resist, thermal lithography including patterning of a spin-coated polymer resist, thermal lithography including patterning of a spin-coated polymer resist, thermal lithography including patterning of a spin-coated polymer resist, the spin-coated polymer resist polymer resist. The spin-coated polymer resist polymer resist$ 

evaporation of a cobalt thin film and lift of f. See Methods section `Sample fabrication' for more details about the individual steps.



Extended Data Fig. 9 | Optical microscope images of a four-panel micromachine, demonstrating operation in water and manipulation of polystyrene microbeads. The micromachine is released in acetone and then a substitution of water for acetone is performed. The magnetization of all four panels points towards the centre.  $\bf a$ , Micromachine in water without a magnetic field.  $\bf b$ , The micromachine panels fold up in an applied out-of-plane magnetic field  $\bf B$  of 10 mT,  $\bf c$ ,  $\bf d$ , On application of a rotating magnetic field  $\bf B$  (10 mT, 5 Hz),

the micromachine rolls across the surface of a silicon wafer, and the rolling motion generates a vortex in the water surrounding it (highlighted with blue arrows). Polystyrene microbeads of 6  $\mu m$  diameter (highlighted with red arrows) are trapped in the vortex and are transported to a new location. Two snapshots of the motion, separated by a time interval of 14 seconds, are shown. Scale bars,  $40\,\mu m$ .



Extended Data Fig. 10 | Optical microscope images of single-panel micromachines operating in air. Two rows of single-panel micromachines are  $shown, each \, suspended \, within \, a \, D\text{-}shaped \, `cutout' \, in \, the \, silicon \, nitride$ membrane frame, and connected to it on the left side by hinge springs with two turns. Each panel is  $10 \, \mu \text{m} \times 10 \, \mu \text{m}$  in size. **a**, After fabrication, the panels are somewhat out-of-focus. This is because they are slightly tilted above the plane of the in-focus silicon nitride frame, which may be due to the residual stress in  $the \ hinge \ springs. \ The \ white \ arrows \ indicate \ the \ magnetization \ direction \ of \ the$ single-panel micromachines pointing left (top row) and right (bottom row). **b**, **c**, On slowly increasing the applied out-of-plane magnetic field, the panels tilt downwards (top row) or upwards (bottom row), with the tilt angle increasing as the field magnitude is increased. In the optical images, the panels with  $magnetization\ pointing\ to\ the\ left\ (top\ row)\ first\ become\ sharper\ ({\bm b}), and\ they$ almost disappear at a tilt angle close to  $90^{\circ}$  at  $35\,\text{mT}$  (c). The panels with magnetization direction pointing to the right (bottom row) tilt upwards in the applied magnetic field, becoming less visible as the field is increased (  $\boldsymbol{b}$  ), until finally disappearing when the tilt angle is close to  $90^{\circ}$  at 35 mT (c). Scale bar (for a-c), 20  $\mu$ m.

# Dry double-sided tape for adhesion of wet tissues and devices

https://doi.org/10.1038/s41586-019-1710-5

Received: 17 September 2018

Accepted: 20 August 2019

Published online: 30 October 2019

Hyunwoo Yuk<sup>1,8</sup>, Claudia E. Varela<sup>2,3,8</sup>, Christoph S. Nabzdyk<sup>4,5</sup>, Xinyu Mao<sup>1</sup>, Robert F. Padera<sup>6</sup>, Ellen T. Roche<sup>1,2,3</sup> & Xuanhe Zhao<sup>1,7\*</sup>

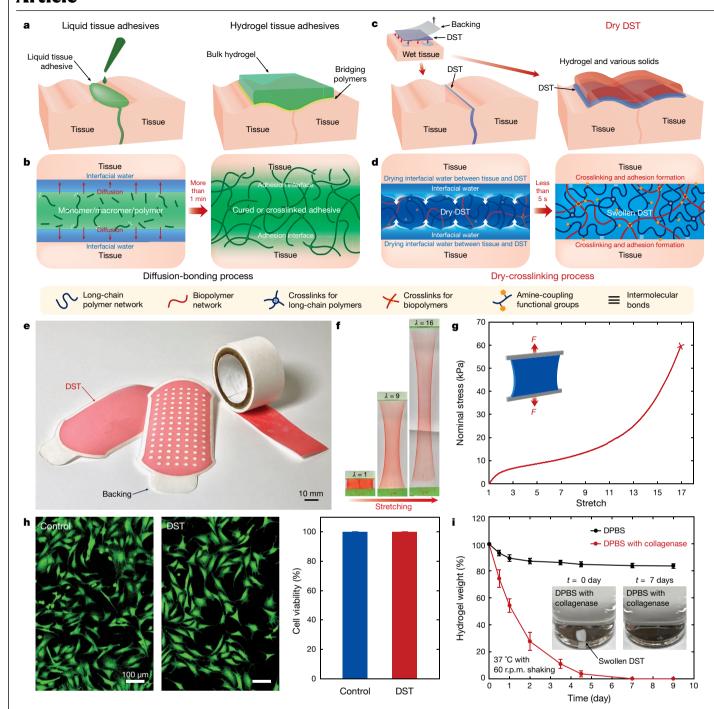
Two dry surfaces can instantly adhere upon contact with each other through intermolecular forces such as hydrogen bonds, electrostatic interactions and van der Waals interactions<sup>1,2</sup>. However, such instant adhesion is challenging when wet surfaces such as body tissues are involved, because water separates the molecules of the two surfaces, preventing interactions<sup>3,4</sup>. Although tissue adhesives have potential advantages over suturing or stapling<sup>5,6</sup>, existing liquid or hydrogel tissue adhesives suffer from several limitations: weak bonding, low biological compatibility, poor  $mechanical\ match\ with\ tissues, and\ slow\ adhesion\ formation^{5-13}.\ Here\ we\ propose\ an$ alternative tissue adhesive in the form of a dry double-sided tape (DST) made from a combination of a biopolymer (gelatin or chitosan) and crosslinked poly(acrylic acid) grafted with N-hydrosuccinimide ester. The adhesion mechanism of this DST relies on the removal of interfacial water from the tissue surface, resulting in fast temporary crosslinking to the surface. Subsequent covalent crosslinking with amine groups on the tissue surface further improves the adhesion stability and strength of the DST. In vitro mouse, in vivo rat and ex vivo porcine models show that the DST can achieve strong adhesion between diverse wet dynamic tissues and engineering solids within five seconds. The DST may be useful as a tissue adhesive and sealant, and in adhering wearable and implantable devices to wet tissues.

Existing tissue adhesives—which are in the form of liquids or wet hydrogels – mostly rely on the diffusion of their molecules (for example, monomers, macromers or polymers) through the interfacial water to form bonds with the polymer networks of tissues<sup>5-12</sup> (Fig. 1a, b). By contrast, animals capable of forming adhesion in wet environments commonly possess mechanisms (for example, mussel, barnacle and spider-web glues) to remove interfacial water from the contact surfaces in order to form bonds<sup>14–16</sup>. Inspired by these examples in nature, we have engineered our DST to adopt a dry-crosslinking mechanism to remove interfacial water and form adhesion on wet tissues (Fig. 1c, d and Extended Data Fig. 1).

The DST consists of two major components: first, poly(acrylic acid) grafted with N-hydroxysuccinimide ester (PAAc-NHS ester) crosslinked by biodegradable gelatin methacrylate; and second, biodegradable biopolymers (for example, gelatin or chitosan). The negatively charged carboxylic acid groups in the PAAc-NHS ester facilitate the quick hydration and swelling of the DST to dry the wet surfaces of various tissues under gentle pressure of around 1 kPa, applied for less than 5 seconds. (See Supplementary Information for a quantitative model showing how the DST dries interfacial water, and Supplementary Figs. 1-6.) Simultaneously, the carboxylic acid groups in the PAAc-NHS ester form intermolecular bonds (for example, hydrogen bonds and electrostatic interactions) with the tissue surfaces (Fig. 1d and Extended Data Fig. 1). To provide stable adhesion, the NHS ester groups grafted on the PAAc also couple covalently with primary amine groups on various tissues within a few minutes, without the need for further pressure (Extended Data Figs. 1.2). After adhering onto tissues, the swollen DST becomes a thin hydrogel layer with an equilibrium water content of around 92% by volume (Extended Data Fig. 3). Because the swollen DST integrates mechanisms for high stretchability and mechanical dissipation<sup>17,18</sup>, it exhibits a high fracture toughness of more than 1,000 J m<sup>-2</sup>  $(Supplementary Figs.\,7,8), which is {\it crucial in achieving tough adhesion}$ of the swollen DST<sup>19,20</sup>.

The dry DST takes the form of a conformable thin film that can be applied on non-planar tissue surfaces (Extended Data Fig. 4). It can be fabricated into diverse shapes such as flat sheets, perforated sheets and adhesive-tape-like rolls (Fig. 1e). The DST, in its fully swollen state, exhibits a shear modulus of 2.5-5 kPa and the ability to stretch to more than 16 times the original length, capable of mechanically matching soft tissues<sup>21</sup> (Fig. 1f, g). To remove potentially cytotoxic residual reagents, we thoroughly purified the DST during its preparation (Extended Data Fig. 5). The in vitro biocompatibility of the DST-conditioned medium is comparable to that of the control medium, showing no observable decrease in the in vitro viability of mouse embryonic fibroblasts (MEFs) after 24-h culture (Fig. 1h). The crosslinkers (that is, gelatin methacrylate) for PAAc-NHS ester and the biopolymers (that is, gelatin

Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. 2Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, USA. 3 Harvard-MIT Program in Health Sciences and Technology, Cambridge, MA, USA. 4 Department of Anesthesiology and Perioperative Medicine, Mayo Clinic, Rochester, MN, USA, 5 Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA, 6 Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. <sup>8</sup>These authors contributed equally: Hyunwoo Yuk, Claudia E, Varela, \*e-mail: zhaox@mit.edu



**Fig. 1**| **Dry DST and dry-crosslinking mechanism for adhesion of wet tissues and devices. a**, Existing tissue adhesives take the form of liquids or wet hydrogels. **b**, Adhesion formation by these existing adhesives mostly relies on the diffusion of monomers, macromers or polymers towards the tissues. **c**, Our proposed tissue adhesive takes the form of a dry DST. **d**, The dry-crosslinking mechanism for the DST integrates the drying of interfacial water by hydration and swelling of the dry DST, temporary crosslinking, and covalent crosslinking (the latter involving the formation of covalent bonds between the DST and amine groups on tissues). **e**, The DST can take on various shapes owing to its

high flexibility in fabrication. The DST is coloured with a red food dye for visualization.  ${\bf f}, {\bf g}$ , Photographs ( ${\bf f}$ ) and nominal stress versus stretch curve ( ${\bf g}$ ) for the DST in its swollen state, stretched to more than 16 times the original length. The DST is coloured with a red food dye for visualization.  $\lambda$ , stretch; F, force.  ${\bf h}$ , In vitro biocompatibility of the DST in a live/dead assay of mouse embryonic fibroblasts (MEFs) after 24 hours of culture.  ${\bf i}$ , In vitro biodegradation of the gelatin-based DST in Dulbecco's phosphate-buffered saline (DPBS) with or without collagenase. Values in panels  ${\bf h}$ ,  ${\bf i}$  represent the mean and the standard deviation (n=3-5).

or chitosan) in the DST are biodegradable by endogenous enzymes (for example, collagenase, lysozyme or N-acetyl- $\beta$ -D-glucosaminidase (NAGase)) at varying rates. For example, gelatin typically degrades more quickly than chitosan under physiological conditions<sup>22</sup>. Hence, the in vitro biodegradation rate of the DST can be controlled over time periods from a week (for the gelatin-based DST) to several months

(for the chitosan-based DST) by tuning its composition (Fig. 1i and Supplementary Fig. 9).

To evaluate the adhesion performance of the DST, we conduct three different types of mechanical tests, measuring the interfacial toughness by peel tests, the shear strength by lap-shear tests and the tensile strength by tensile tests (according to the following testing standards

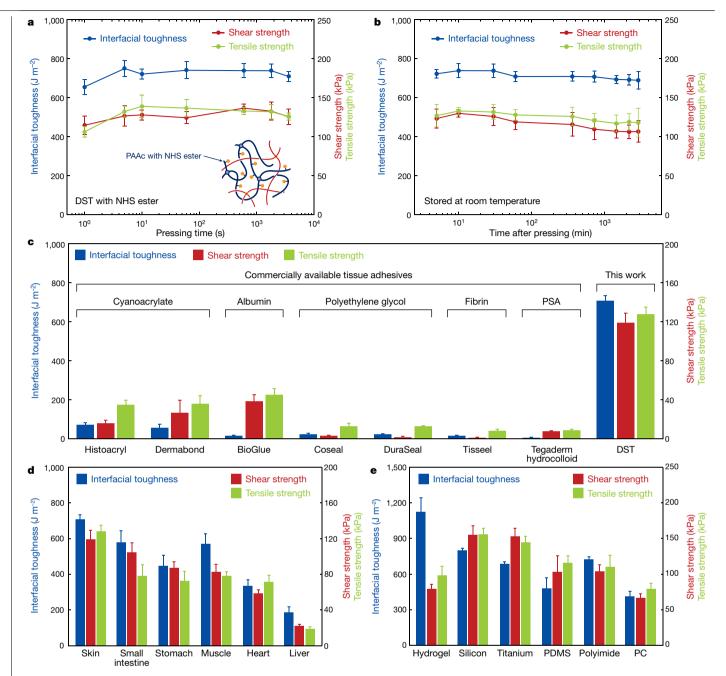


Fig. 2 | Adhesion performance of the DST. a, Interfacial toughness and shear and tensile strength versus pressing time for wet porcine skins adhered using the DST with NHS ester. b, Interfacial toughness and shear and tensile strength versus time after pressing for wet porcine skins adhered using the DST with NHS ester. c, Comparison of adhesion performances of the DST and various

commercially available tissue adhesives. PSA, pressure-sensitive adhesives. d, Interfacial toughness and shear and tensile strength between various tissues adhered by the DST. e, Interfacial toughness and shear and tensile strength between porcine skin and various engineering solids adhered by the DST. Values represent the mean and standard deviation (n = 3-5).

for tissue adhesives: ASTM F2256 for 180-degree peel tests, ASTM F2255 for lap-shear tests, and ASTM F2258 for tensile tests; Extended Data Fig. 6 and Supplementary Fig. 10). We first choose wet porcine skin as the model tissue for evaluation of adhesion performance, owing to its mechanical robustness and close resemblance to human skin<sup>12</sup>. The DST can establish tough (with an interfacial toughness of more than 710 J m<sup>-2</sup>) and strong (with a shear and tensile strength of more than 120 kPa) adhesion between wet porcine skins upon contact and application of gentle pressure (1 kPa) for less than 5 seconds (Fig. 2a, Supplementary Fig. 11 and Supplementary Video 1). The tissues adhered by the DST exhibit a relatively small decrease (of less than 10%) in the measured interfacial toughness and strength more than 48 h after the initial pressing (Fig. 2b). Furthermore, the DST can maintain its ability

to form robust adhesion on wet tissues after being stored for more than 2 weeks (Supplementary Fig. 12).

We also examine the importance of covalent crosslinking after intermolecular crosslinking on the adhesion performance of the DST. We test the adhesion performance of the DST without grafted NHS ester on the PAAc, which cannot form covalent crosslinks with tissues (Extended Data Fig. 7). Although the DST without NHS ester can provide tough (with an interfacial toughness greater than 500 J m<sup>-2</sup>) and strong (with a shear and tensile strength more than 80 kPa) adhesion upon application between wet porcine skins (Extended Data Fig. 7a), the adhesion performance shows substantial deterioration over time (Extended Data Fig. 7b), owing to the unstable and temporary nature of the intermolecular bonds in wet environments<sup>3</sup>. Hence, both the

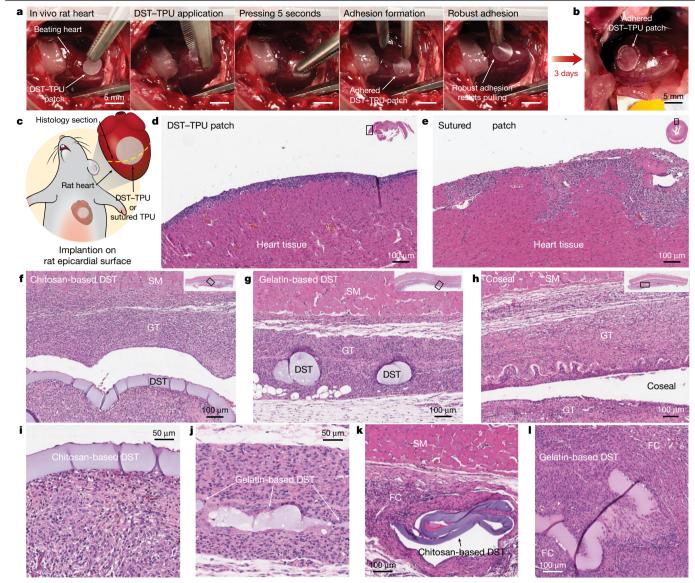


Fig. 3| In vivo adhesion, biocompatibility, and biodegradability of the DST. a, Adhesion of a DST–TPU hybrid patch on a beating rat heart in vivo. b, The DST–TPU patch adhered on the rat heart 3 days after implantation in vivo. c, A schematic illustration of the section taken for histology (dotted yellow line) through a DST–TPU or a sutured TPU patch implanted on the rat epicardial surface. d, e, Representative histological images of the DST–TPU patch (d) and the sutured TPU patch (e) stained with haematoxylin and eosin (H&E). f–h, Representative histological images of the chitosan-based DST (f), the

gelatin-based DST(g) and the Coseal(h), stained with H&E.i, j, Representative histological images stained with H&E for assessment of the biodegradation of subcutaneously implanted chitosan-based DST(i) and gelatin-based DST(j) after 2 weeks. k, l, Representative histological images stained with H&E for assessment of the biocompatibility and biodegradation of subcutaneously implanted chitosan-based DST(k) and gelatin-based DST(l) after 4 weeks. SM, GT and FC indicate skeletal muscle, granulation tissue and fibrous capsule, respectively. All experiments were repeated three or four times with similar results.

temporary crosslinks and the subsequent covalent crosslinks are necessary for stable and robust adhesion of wet surfaces, supporting our proposed mechanism (Fig. 1d and Extended Data Fig. 1).

We further test the adhesion performance of the DST under cyclic loading conditions. Two porcine heart tissues adhered by the DST maintain a high interfacial toughness of more than 650 J m<sup>-2</sup> during cyclic loading over 5,000 cycles with physiologically relevant strain (30% tensile strain) (Supplementary Fig. 13). In addition, the DST can provide similarly high interfacial toughness (of more than 640 J m<sup>-2</sup>) and shear and tensile strength (of more than 85 kPa) on blood-covered porcine tissues after washout with saline<sup>23</sup> (Supplementary Fig. 14).

The DST demonstrates superior adhesion performance compared with existing tissue adhesives, including commercially available cyanoacrylate adhesives (Histoacryl and Dermabond), albumin-based adhesives (BioGlue), polyethylene-glycol-based adhesives (Coseal and

DuraSeal), fibrin glues (Tisseel), hydrophilic pressure-sensitive adhesives (Tegaderm hydrocolloid), as well as nanoparticle solutions and ultraviolet-curable surgical glues  $^{24}$  (Fig. 2c and Extended Data Fig. 8). We find that these existing tissue adhesives require a relatively long time to form adhesion (longer than 1 min; Extended Data Fig. 8) and exhibit limited adhesion performance on wet tissues (with an interfacial toughness of less than than 20 J m $^{-2}$  and a shear and tensile strength of less than 45 kPa; Fig. 2c), consistent with published performances  $^{7}$ . The DST provides a higher interfacial toughness (up to 1,150 J m $^{-2}$ ) and shear and tensile strength (up to 160 kPa) than existing tissue adhesives, and forms the adhesion in less than 5 seconds (Fig. 2c–e and Extended Data Fig. 8). Although tough hydrogel adhesives can achieve a similarly high interfacial toughness of more than 1,000 J m $^{-2}$  on wet tissues, they require steady pressure application for substantially longer periods of time (5–30 min) on tissue surfaces to form the adhesion  $^{12,25}$ .

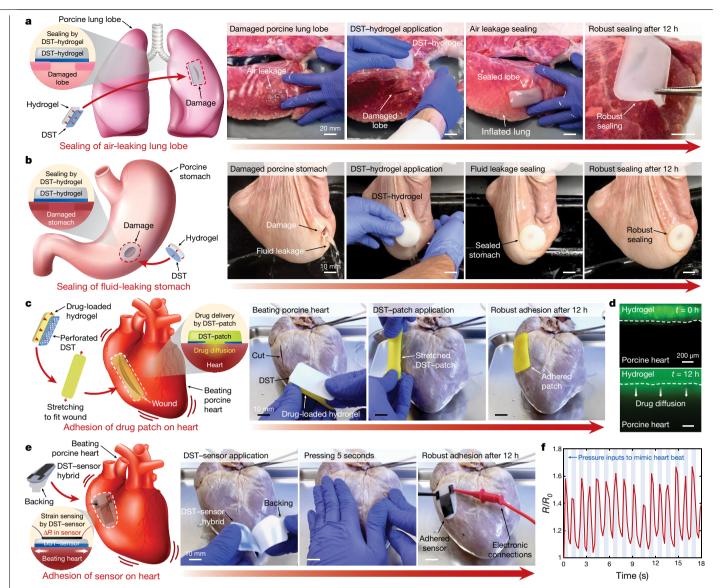


Fig. 4 | Potential applications of the DST. a, Sealing of an air-leaking lacerated ex vivo porcine lung lobe by a hydrogel patch adhered with the DST. b, Sealing of a fluid-leaking ex vivo porcine stomach by a hydrogel patch adhered with the DST. c, DST-mediated adhesion of a drug-loaded patch on a beating ex vivo porcine heart with a cut. d, Diffusion of a mock drug (fluorescein) from a DSTadhered drug patch into the ex vivo porcine heart tissue over time. e, Adhesion

of a DST-strain-sensor hybrid on a beating ex vivo porcine heart. f, Normalized electrical resistance  $(R/R_0)$  of the DST-adhered strain sensor over time, in order to measure deformation of the beating heart. The blue shades in the graph indicate the intervals during which pressure inputs are introduced to the ex vivo porcine heart to mimic beating.

The DST can be applied to various wet tissues, including skin, small intestine, stomach, muscle, heart and liver (Fig. 2d and Supplementary Video 2), with high interfacial toughness (more than 710 J m<sup>-2</sup> for skin,  $580 \, \mathrm{J} \, \mathrm{m}^{-2}$  for small intestine,  $450 \, \mathrm{J} \, \mathrm{m}^{-2}$  for stomach,  $570 \, \mathrm{J} \, \mathrm{m}^{-2}$  for muscle, 340 J m<sup>-2</sup> for heart and 190 J m<sup>-2</sup> for liver) and high shear and tensile strength (more than 120 kPa for skin, 80 kPa for small intestine, 70 kPa for stomach, 80 kPa for muscle, 70 kPa for heart and 20 kPa for liver) (Fig. 2d and Supplementary Fig. 10a-c). The DST can also provide adhesion between wet tissues and various engineering solids, including hydrogel, silicon, titanium, polydimethylsiloxane (PDMS), polyimide and polycarbonate (Fig. 2e and Supplementary Video 3). We functionalize the surfaces of various engineering solids with primary amines in order to ensure covalent coupling with the DST<sup>19</sup> (Extended Data Fig. 9), and then evaluate the adhesion performance using wet porcine skin (Supplementary Fig. 10d-f). The adhesion between the wet tissues and various engineering solids by the DST exhibits high interfacial toughness (higher than 1,150 J m<sup>-2</sup> for hydrogel, 800 J m<sup>-2</sup> for silicon, 680 J m<sup>-2</sup> for titanium, 480 J m<sup>-2</sup> for PDMS, 720 J m<sup>-2</sup> for polyimide and 410 J m<sup>-2</sup> for

polycarbonate) and high shear and tensile strength (more than 80 kPa for hydrogel, 160 kPa for silicon, 150 kPa for titanium, 100 kPa for PDMS, 100 kPa for polyimide and 70 kPa for polycarbonate) (Fig. 2e).

In order to evaluate the ability of the DST to adhere to wet and dynamic surfaces in vivo, we adhere a thermoplastic polyurethane (TPU) patch<sup>26</sup> to the epicardial surface of a rat heart using the DST (Fig. 3a, b and Supplementary Fig. 15a). We find that a 5-mm-diameter DST-TPU hybrid patch (using the gelatin-based DST with a dry thickness of 20 µm) can be adhered to the epicardial surface of a beating rat heart after gently pressing for 5 seconds (Fig. 3a and Supplementary Video 4). After 3 days of in vivo implantation, the DST-TPU patch maintains adhesion to the rat heart surface (Fig. 3b) while producing a host response similar to that of reported epicardial patches (Fig. 3c, d). Histological assessment by a blinded pathologist indicates that the degree of inflammatory reaction induced by the DST-TPU patch is comparable to that of a sutured TPU patch (Fig. 3c-e).

We further evaluate the in vivo biocompatibility and biodegradability of the DST in a rat model of dorsal subcutaneous implantation

(Fig. 3f-h and Supplementary Fig. 15b). Histological assessment demonstrates that, after 2 weeks of implantation, the chitosan-based DST (20-µm dry thickness) generates a comparable inflammatory reaction (Fig. 3f) to that produced by Coseal, a US Food and Drug Administration (FDA)-approved, commercially available tissue adhesive (Fig. 3h). The histology at 2 weeks for both implant types is characterized by a mild to moderate chronic inflammatory response involving macrophages, lymphocytes and occasional giant cells in association with the formation of a capsule of granulation tissue comprising fibroblasts, collagen and new blood vessels. There is no evidence of necrosis of the overlying skeletal muscle or skin, or of an eosinophilic response suggestive of an allergic reaction. Although the gelatin-based DST shows a higher degree of inflammatory response than the chitosan-based DST, as indicated by a denser chronic inflammatory reaction (Fig. 3g), no major damage to the surrounding dermal and muscular layers is observed after 2 weeks of implantation. The more pronounced inflammatory response of the gelatin-based DST might result from the faster biodegradation of gelatin than of chitosan and subsequent effects on the surrounding tissues, including a higher degree of phagocytotic responses<sup>27</sup>.

Furthermore, the histological images of the subcutaneously implanted DST demonstrate in vivo biodegradability of the DST (Fig. 3i–l). After two weeks of implantation, the relatively slow-degrading chitosan-based DST maintains an intact film-like configuration (Fig. 3i), while the relatively fast-degrading gelatin-based DST shows signs of degradation such as reduction in thickness and a fragmented configuration (Fig. 3j). At an implantation period of 4 weeks, the chitosan-based DST shows limited signs of degradation (Fig. 3k), whereas there is substantial continued degradation of the gelatin-based DST, as shown by increased material resorption by macrophages (Fig. 3l). In addition, there is appropriate evolution and attenuation of the inflammatory response generated by both the gelatin-based and the chitosan-based DST after 4 weeks of implantation, including a decrease in the magnitude of the chronic inflammatory infiltrate and thinning of the surrounding fibrous capsule.

In order to demonstrate potential applications of the DST, we investigate a range of proof-of-principle applications using ex vivo porcine models. The DST combined with a degradable tough hydrogel patch can form an air-tight sealing of a lacerated, air-leaking lung lobe and trachea (Fig. 4a, Extended Data Fig. 10a and Supplementary Video 5). Furthermore, it provides a fluid-tight sealing of a fluid-filled perforated stomach (with a 1-cm-wide hole) and a dissected small intestine (Fig. 4c. Extended Data Fig. 10b and Supplementary Videos 6, 7). We further show that the DST can be used to adhere devices onto dynamic and deformable tissues<sup>26,28-30</sup>. For example, we use the DST to adhere a hydrogel patch with a mock drug (fluorescein) onto a beating ex vivo porcine heart (introducing cyclical, pressurized air inputs to mimic heart beats). This suggests that the DST might allow the attachment of drug-delivery devices onto dynamic wet tissues (Fig. 4d and Supplementary Video 8). The adhered DST patch maintains adhesion on the beating heart for more than 12 h without any sign of decreased adhesion and allows delivery of a mock drug into the heart tissue (Fig. 4e). As another example, we adhere a stretchable strain sensor on the beating porcine heart (Fig. 4f and Supplementary Video 9). The DST allows facile attachment of the strain sensor on the dynamic and curved surface of the beating heart, as well as electrical measurements of the heart movements (Fig. 4g). Notably, the stretchable DST-sensor hybrid is prepared by printing a conductive ink on a DST-Ecoflex hybrid substrate<sup>20</sup> (Supplementary Fig. 16). Such DST-device hybrids could serve as a versatile platform for wearable and implantable devices to adhere onto wet and dynamic tissues. Although these ex vivo models show possible applications of the DST, we note that its long-term efficacy, biocompatibility and biodegradability, as well as the induced biological responses (for example, healing) in clinically relevant settings, require further studies.

In this study, we have reported a biologically inspired, dry-crosslinking mechanism—which is implemented in the form of a dry DST—for the

adhesion of wet tissues and devices. This DST offers advantages over existing tissue adhesives and sealants, including fast adhesion formation, robust adhesion performance, flexibility, and ease of storage and use. The DST may also provide new opportunities for bioscaffolds, drug delivery, and wearable and implantable devices. This dry-crosslinking mechanism for the adhesion of wet surfaces could also be applied in the design of adhesives for wet and underwater environments.

#### **Online content**

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1710-5.

- Creton, C. Pressure-sensitive adhesives: an introductory course. MRS Bull. 28, 434–439 (2003).
- 2. Chung, J. Y. & Chaudhury, M. K. Soft and hard adhesion. J. Adhes. 81, 1119-1145 (2005).
- Peppas, N. A. & Buri, P. A. Surface, interfacial and molecular aspects of polymer bioadhesion on soft tissues. J. Control. Release 2, 257–275 (1985).
- Lee, H., Lee, B. P. & Messersmith, P. B. A reversible wet/dry adhesive inspired by mussels and geckos. Nature 448, 338–341 (2007).
- Reece, T. B., Maxey, T. S. & Kron, I. L. A prospectus on tissue adhesives. Am. J. Surg. 182, S40–S44 (2001).
- Coulthard, P. et al. Tissue adhesives for closure of surgical incisions. Cochrane Database Syst. Rev. 11, CD004287 (2014).
- Vakalopoulos, K. A. et al. Mechanical strength and rheological properties of tissue adhesives with regard to colorectal anastomosis: an ex vivo study. Ann. Surg. 261, 323–331 (2015).
- Rose, S. et al. Nanoparticle solutions as adhesives for gels and biological tissues. Nature 505, 382–385 (2014).
- Lee, B. P., Messersmith, P. B., Israelachvili, J. N. & Waite, J. H. Mussel-inspired adhesives and coatings. Annu. Rev. Mater. Res. 41, 99–132 (2011).
- Annabi, N., Yue, K., Tamayol, A. & Khademhosseini, A. Elastic sealants for surgical applications. Eur. J. Pharm. Biopharm. 95, 27–39 (2015).
- 11. Karp, J. M. A slick and stretchable surgical adhesive. N. Engl. J. Med. 377, 2092–2094 (2017).
  - Karp, J. M. A slick and stretchable surgical adhesive. N. Engl. J. Med. 377, 2092–209
     Li, J. et al. Tough adhesives for diverse wet surfaces. Science 357, 378–381 (2017).
- LeMaire, S. A. et al. The threat of adhesive embolization: BioGlue leaks through needle holes in aortic tissue and prosthetic grafts. Ann. Thorac. Surg. 80, 106–111 (2005).
- Maier, G. P., Rapp, M. V., Waite, J. H., Israelachvili, J. N. & Butler, A. Adaptive synergy between catechol and lysine promotes wet adhesion by surface salt displacement. Science 349, 628–632 (2015).
- Zhao, Q. et al. Underwater contact adhesion and microarchitecture in polyelectrolyte complexes actuated by solvent exchange. Nat. Mater. 15, 407–412 (2016).
- Singla, S., Amarpuri, G., Dhopatkar, N., Blackledge, T. A. & Dhinojwala, A. Hygroscopic compounds in spider aggregate glue remove interfacial water to maintain adhesion in humid conditions. *Nat. Commun.* 9, 1890 (2018).
- Gong, J. P., Katsuyama, Y., Kurokawa, T. & Osada, Y. Double-network hydrogels with extremely high mechanical strength. Adv. Mater. 15, 1155–1158 (2003).
- Zhao, X. Multi-scale multi-mechanism design of tough hydrogels: building dissipation into stretchy networks. Soft Matter 10, 672–687 (2014).
- Yuk, H., Zhang, T., Lin, S., Parada, G. A. & Zhao, X. Tough bonding of hydrogels to diverse non-porous surfaces. Nat. Mater. 15, 190–196 (2016).
- Yuk, H., Zhang, T., Parada, G. A., Liu, X. & Zhao, X. Skin-inspired hydrogel-elastomer hybrids with robust interfaces and functional microstructures. Nat. Commun. 7, 12028 (2016).
- 21. Yuk, H., Lu, B. & Zhao, X. Hydrogel bioelectronics. Chem. Soc. Rev. 48, 1642–1667 (2019).
- Gorgieva, S. & Kokol, V. Preparation, characterization, and in vitro enzymatic degradation of chitosan-gelatine hydrogel scaffolds as potential biomaterials. J. Biomed. Mater. Res. A 100A, 1655–1667 (2012).
- Barnes, S., Spencer, M., Graham, D. & Johnson, H. B. Surgical wound irrigation: a call for evidence-based standardization of practice. Am. J. Infect. Control 42, 525–529 (2014).
- Lang, N. et al. A blood-resistant surgical glue for minimally invasive repair of vessels and heart defects. Sci. Transl. Med. 6, 218ra6 (2014).
- Yang, J., Bai, R. & Suo, Z. Topological adhesion of wet materials. Adv. Mater. 30, 1800671 (2018).
- Whyte, W. et al. Sustained release of targeted cardiac therapy with a replenishable implanted epicardial reservoir. Nat. Biomed. Eng. 2, 416–428 (2018).
- Jiang, W. W., Su, S. H., Eberhart, R. C. & Tang, L. Phagocyte responses to degradable polymers. J. Biomed. Mater. Res. A 82A, 492–497 (2007).
- 28. Yamagishi, K. et al. Tissue-adhesive wirelessly powered optoelectronic device for metronomic photodynamic cancer therapy. *Nat. Biomed. Eng.* **3**, 27–36 (2019).
- Feiner, R. et al. Engineered hybrid cardiac patches with multifunctional electronics for online monitoring and regulation of tissue function. Nat. Mater. 15, 679–685 (2016).
- Roche, E. T. et al. Soft robotic sleeve supports heart function. Sci. Transl. Med. 9, eaaf3925 (2017).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

@ The Author(s), under exclusive licence to Springer Nature Limited 2019

#### **Methods**

#### **Materials**

All chemicals were obtained from Sigma-Aldrich unless otherwise mentioned, and used without further purification. To prepare the DST, we used acrylic acid, gelatin methacrylate (gelMA; type A bloom 90-100 from porcine skin with 60% substitution), acrylic acid N-hydroxysuccinimide ester (AAc-NHS ester), α-ketoglutaric acid, gelatin (type A bloom 300 from porcine skin) and chitosan (75–85% deacetylated). To visualize the DST in photographs and microscope images, we used red food dve (McCormick) and fluorescein isothiocvanate (FITC)-gelatin (Thermo Fisher Scientific). For in vitro biodegradation tests, we used Dulbecco's phosphate-buffered saline (DPBS, with calcium and magnesium: Gibco), collagenase, lysozyme and NAGase. To prepare the degradable tough hydrogel, we used acrylamide, gelatin, gelMA, and Irgacure 2959. For surface functionalization of engineering solids, we used (3-aminopropyl) triethoxysilane (APTES) and hexamethyldiamine (HMDA). To prepare the stretchable strain sensor, we used Ecoflex 00-30 (Smooth-On), silicone curing retardant (SLO-JO, Smooth-On) and carbon black (Alfa Aesar). All engineering solids were obtained from McMaster Carr unless otherwise mentioned. All porcine tissues for ex vivo experiments were purchased from a research-grade porcine tissue vendor (Sierra Medical).

#### Preparation of the dry DST

The DST was prepared with either gelatin or chitosan. To prepare the gelatin-based DST, we dissolved 30% (w/w) acrylic acid, 10% (w/w) gelatin, 1% (w/w) AAc-NHS ester, 0.1% (w/w) gelMA and 0.2% (w/w)  $\alpha$ -ketoglutaric acid in deionized water. The mixture was then filtered with 0.4-µm sterile syringe filters and poured on a glass mould with spacers. The DST was cured in an ultraviolet light (UV) chamber (284 nm, 10 W power) for 20 min and completely dried. The final DST was sealed in plastic bags with desiccant (silica gel packets) and stored at -20 °C before use. The chitosan-based DST was prepared by replacing 10% (w/w) gelatin with 2% (w/w) chitosan. In experiments, we used the gelatin-based DST with an as-prepared thickness of 210 µm unless otherwise mentioned. To prepare the DST in various shapes, we cut a large sheet of DST into each design using a laser cutter (Epilog). Polyethylene-coated paper was used as a backing for the DST. To aid visualization of the DST, we added 0.5% (w/w) of red food dye (for photographs) or 0.2% (w/w) FITC-gelatin (for fluorescence microscopy images) into the precursor solution before curing.

#### **Mechanical tests**

Tissue samples stored more than 10 min before mechanical tests were covered with an excess of 0.01% (w/v) sodium azide solution (in PBS) spray and sealed in plastic bags to prevent degradation and dehydration. Unless otherwise indicated, all tissues and engineering solids were adhered by the DST after washing out the surfaces with PBS followed by 5 seconds of pressing (with 1 kPa pressure applied by either a mechanical testing machine or an equivalent weight). Unless otherwise indicated, all mechanical tests on adhesion samples were performed 24 h after initial pressing to ensure equilibrium swelling of the adhered DST in wet environments. The application of commercially available tissue adhesives followed the manual provided for each product. The gelatin-based DST was used unless otherwise noted.

To measure interfacial toughness, adhered samples with widths of 2.5 cm were prepared and tested by the standard 180-degree peel test (ASTM F2256) or 90-degree peel test (ASTM D2861) (for inflexible substrates such as silicon) using a mechanical testing machine (2.5 kN load-cell, Zwick/Roell Z2.5). All tests were conducted with a constant peeling speed of 50 mm min<sup>-1</sup>. The measured force reached a plateau as the peeling process entered the steady state. Interfacial toughness was determined by dividing two times the plateau force (for a 180-degree peel test) or the plateau force (for a 90-degree peel test)

by the width of the tissue sample (Extended Data Fig. 6a). Poly(methyl methacrylate) films (with a thickness of 50  $\mu$ m; Goodfellow) were applied using cyanoacrylate glue (Krazy Glue) as a stiff backing for the tissues and hydrogels.

To measure shear strength, adhered samples with an adhesion area of width 2.5 cm and length 1 cm were prepared and tested by the standard lap-shear test (ASTM F2255) with a mechanical testing machine (2.5 kN load-cell, Zwick/Roell Z2.5) (Extended Data Fig. 6b). All tests were conducted with a constant tensile speed of 50 mm min $^{-1}$ . Shear strength was determined by dividing the maximum force by the adhesion area. Poly(methyl methacrylate) films were applied using cyanoacrylate glue to act as a stiff backing for the tissues and hydrogels.

To measure tensile strength, adhered samples with adhesion areas of width 2.5 cm and length 2.5 cm were prepared and tested by the standard tensile test (ASTM F2258) with a mechanical testing machine (2.5 kN load-cell, Zwick/Roell Z2.5) (Extended Data Fig. 6c). All tests were conducted with a constant tensile speed of 50 mm min $^{-1}$ . Tensile strength was determined by dividing the maximum force by the adhesion area. Aluminium fixtures were applied using cyanoacrylate glues to provide grips for tensile tests.

To characterize mechanical properties of the swollen DST, the DST was equilibrated in PBS before tests. The tensile properties and fracture toughness of the DST were measured using pure-shear tensile tests of thin rectangular samples (10 mm in length, 30 mm in width and 0.5 mm in thickness) with a mechanical testing machine (20 Nload-cell, Zwick/Roell Z2.5). All tests were conducted with a constant tensile speed of 50 mm min<sup>-1</sup>. The fracture toughness of the DST was calculated using a reported method based on tensile tests of unnotched and notched samples<sup>12</sup>.

To characterize the adhesion performance of the DST under cyclic loading, two porcine heart tissues were adhered by the DST with an adhesion area of width 2.5 cm and length 4 cm. Each side of the adhered tissues was cyclically stretched at 30% tensile strain (with respect to the DST length) using a mechanical testing machine (2.5 kN load-cell, Zwick/Roell Z2.5) to provide cyclic shear loading to the adhesion interface (Supplementary Fig. 13). Interfacial toughness between the heart tissues adhered by the DST was measured at different cycle numbers by the standard 180-degree peel test (ASTM F2256). During the cyclic tests, a 0.01% (w/v) sodium azide solution (in PBS) was sprayed onto the heart tissues to avoid tissue degradation and dehydration.

#### **Preparation of engineering solids**

To prepare degradable tough hydrogels for adhesion tests of engineering solids, we dissolved 20% (w/w) acrylamide, 10% (w/w) gelatin, 0.2% (w/w) gelMA and 0.2% (w/w) Irgacure 2959 in deionized water. The mixture was then filtered with 0.4-µm sterile syringe filters and poured on a glass mould with spacers. The hydrogels were cured in a UV chamber (284 nm, 10 W power) for 60 min. To facilitate covalent coupling with the DST, engineering solids except hydrogel were functionalized with primary amines (Extended Data Fig. 9). For silicon, titanium and PDMS, the substrates were first treated with oxygen plasma for 2 min (30 W power, Harrick Plasma) to activate the surface. Subsequently, the plasma-treated substrates were covered with APTES solution (1% (w/w) APTES in 50% ethanol) and incubated for 3 h at room temperature<sup>20</sup>. The substrates were then thoroughly washed with isopropyl alcohol and dried using a nitrogen flow. For polyimide and polycarbonate, the substrates were immersed into the HMDA solution (10% (v/v) in deionized water) for 24 hat room temperature. The substrates were then thoroughly washed with deionized water and dried using a nitrogen flow<sup>31</sup>.

#### In vitro biocompatibility tests

We conducted in vitro biocompatibility tests using DST-conditioned medium for cell culture  $^{32}$ . To prepare the DST-conditioned medium for in vitro biocompatibility tests, we incubated 20 mg of the gelatin-based DST in 1 ml of Dulbecco's modified Eagle medium (DMEM) at 37  $^{\circ}\mathrm{C}$ 

for 24 h. Pristine DMEM was used as a control. Wild-type MEFs were plated in 96-well plates ( $n\!=\!10$  for DST-conditioned medium;  $n\!=\!10$  for DMEM). The cells were then treated with the DST-conditioned medium and incubated at 37 °C for 24 h in 5% CO2. Cell viability was determined with a live/dead viability/cytotoxicity kit for mammalian cells (Thermo Fisher Scientific) by adding 4  $\mu$ M calcein and ethidium homodimer-1 into the culture medium. We used a confocal microscope (SP 8, Leica) to image live cells with excitation/emission at 495 nm/515 nm, and dead cells at 495 nm/635 nm.

#### In vitro biodegradation tests

We carried out in vitro biodegradation tests of the DST using enzymatic degradation media as described<sup>22</sup>. To prepare in vitro enzymatic biodegradation medium for the gelatin-based DST, we added 5 mg collagenase in 100 ml DPBS. To prepare in vitro enzymatic biodegradation medium for the chitosan-based DST, we added 5 mg collagenase, 5 mg lysozyme and 10 µl of 1 mg ml<sup>-1</sup> NAGase aqueous solution in 100 ml DPBS. The DST was cut into small samples (of width 10 mm and length 10 mm) and accurately weighed. Before immersion in the enzymatic media, the samples were sterilized in 75% ethanol for 15 min and washed three times with DPBS. Each sample was then immersed in 15 ml of the enzymatic medium within glass scintillation vials and incubated at 37 °C with shaking at 60 r.p.m. About 0.01% (w/v) sodium azide was added into the enzymatic media to prevent growth of any microorganisms during the tests. At each time interval, the DST was removed from the incubation medium, exhaustively washed with deionized water and lyophilized. Weight loss was determined as the percentage ratio of the mass of the lyophilized sample at each time interval, normalized by the dry mass of the original lyophilized sample.

#### In vivo adhesion, biocompatibility and biodegradability tests

All animal procedures were reviewed and approved by the Massachusetts Institute of Technology Committee on Animal Care. Details of surgical procedures and in vivo data analysis are provided in the Supplementary Information.

#### Preparation of the DST-strain-sensor hybrid

We prepared the DST-strain-sensor hybrid by printing a conductive ink onto a DST-elastomer hybrid substrate. This elastomer substrate was first prepared by casting Ecoflex 00-30 resin mixture (part A and part B in a 1/1 volume ratio) into a laser-cut acrylic mould. Subsequently, a thin layer of gelatin-based DST (100-µm dry thickness) was introduced on the bottom side of the Ecoflex substrate according to the reported protocol for hydrogel-elastomer hybrids<sup>21</sup>. The strain sensor was fabricated by printing the conductive ink onto the DST-Ecoflex hybrid substrate using a custom direct ink writing (DIW) 3D printer<sup>33</sup>. Briefly, the conductive ink was prepared by mixing 10% (w/w) carbon black and 1% (w/w) silicone curing retardant into Ecoflex 00-30 resin (part A and part B in a 1/1 volume ratio) using a planetary mixer (AR-100, Thinky). The printing paths were generated through production of G-codes that control the XYZ motions of a robotic gantry (Aerotech). We used a pressure-based microdispenser (Ultimus V, Nordson EFD) with a 200-µm-diameter nozzle (Smoothflow tapered tip, Nordson EFD) to print the conductive ink on the substrate through a custom LabVIEW interface (National Instruments). Deformation-induced changes in the electrical resistance of the strain sensor were monitored with a digital multimeter (34450A, Keysight).

#### **HPLC** characterization of the DST

We analysed the residual monomer contents of the DST using analytical high-performance liquid chromatography (HPLC; Model 1100, Agilent). We used 0.1% phosphoric acid as the mobile phase, extractant and medium for an acrylic acid monomer standard solution as described  $^{34}$  (Extended Data Fig. 5). To extract the residual monomer from the DST, we incubated 100 mg of the DST in 20 ml of the extractant for 24 h with

stirring. After the extraction, the solution was filtered with a sterile 0.2-µm syringe filter and injected into the HPLC system for analysis. The concentration of the residual acrylic acid monomer in the DST was determined on the basis of the calibration curve obtained from the standard solution diluted with the mobile phase to varying monomer concentrations.

#### FTIR characterization of the DST

The chemical composition of the DST was characterized using a transmission Fourier transform infrared spectroscope (FTIR 6700, Thermo Fisher) with a germanium-attenuated total reflectance (ATR) crystal (55 degrees). The FTIR spectrum of the DST was analysed as described (Extended Data Fig. 2a).

#### Ex vivo tests

All ex vivo experiments were reviewed and approved by the Committee on Animal Care at the Massachusetts Institute of Technology. To assess sealing of damaged trachea, we made a laceration (1.5 cm in length) in a porcine trachea using a razor blade. Air was then applied through tubing connected to the upper part of the trachea (25 mm Hg pressure) to visualize air leakage from the trachea submerged in a water bath. To seal the laceration, we adhered a hydrogel patch (of width 2.5 cm and length 5 cm) to the damaged trachea using the DST with 5 seconds of pressing. The sealed porcine trachea was kept for 12 h at room temperature with continuous inflation—deflation cycles to monitor the DST-based sealing. We added 0.01% (w/v) sodium azide into the water bath to avoid tissue degradation.

To assess sealing of a damaged lung lobe, we made a laceration (3 cm long) in a porcine lung lobe with a razor blade. Air was then applied through tubing connected to the upper part of the trachea (25 mm Hg pressure) in order to visualize air leakage from the lung lobe submerged in the water bath. To seal the laceration, we adhered a hydrogel patch (of width 2.5 cm and length 5 cm) to the damaged lung lobe using the DST with 5 seconds of pressing. The sealed porcine lung lobe was kept for 12 h at room temperature with continuous inflation—deflation cycles to monitor the DST-based sealing. We added 0.01% (w/v) sodium azide into the water bath to avoid tissue degradation.

To assess sealing of damaged stomach, we punched a 10-mm-wide hole in a porcine stomach. A tube with flowing water was then connected to the upper part of the stomach to visualize fluid leakage from the stomach. To seal the hole, we adhered a 40-mm-wide hydrogel patch onto the damaged stomach using the DST with 5 seconds of pressing. The sealed porcine stomach was kept for 12 h at room temperature to monitor the DST-based sealing. We sprayed 0.01% (w/v) sodium azide solution (in PBS) onto the porcine stomach to avoid tissue degradation.

To assess sealing of an anastomosis site in a small intestine, we dissected a porcine small intestine into two pieces. Anastomosis of the dissected small intestine was made by approximating each edge of the small intestine followed by wrapping of the DST (2.5 cm wide and 8 cm long) and 5 seconds of pressing around the approximated edges. To check that the DST had produced fluid-tight sealing, we applied water to the anastomosed small intestine at 60 mm Hg pressure using a microdispenser. We sprayed 0.01% (w/v) sodium azide solution (in PBS) onto the porcine small intestine to avoid tissue degradation.

To assess the adhesion of a drug-delivery device, we introduced a cut (4 cm in length) on an explanted porcine heart. The aorta was connected to tubing, and programmed pressurized air inputs were introduced into the heart using a microdispenser to mimic heart beats. To prepare the drug-delivery device, we added 0.5% (w/w) fluorescein sodium salt as a mock drug into a hydrogel patch (2.5 cm in width and 5 cm in length). The drug-loaded hydrogel patch was then stretched to fit the cut and adhered onto the beating porcine heart with the perforated DST. The adhered drug patch on the beating heart was kept for 12 h

at room temperature with continuous beating to allow diffusion of the mock drug into the heart tissue. The diffusion of the mock drug was imaged using a fluorescence microscope (LV100ND, Nikon). To assess the adhesion of a strain sensor, we adhered the DST–strain-sensor hybrid onto the beating porcine heart after removing the backing. The adhered strain sensor on the beating heart was kept for 12 h at room temperature with continuous beating, and then connected with the digital multimeter to monitor the deformation of the beating heart. All devices were adhered onto the beating heart after washing out the surfaces with PBS, followed by 5 seconds of pressing. To prevent dehydration and degradation during experiments of longer than 1 h in ambient conditions, we covered the heart with a wet towel soaked with 0.01% (w/v) sodium azide solution (in PBS).

#### **Reporting summary**

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

#### **Data availability**

 $All\,data\,are\,available\,in\,the\,main\,text\,or\,the\,Supplementary\,Information.$ 

- VanDelinder, V. et al. Simple, benign, aqueous-based amination of polycarbonate surfaces. ACS Appl. Mater. Interfaces 7, 5643–5649 (2015).
- Darnell, M. C. et al. Performance and biocompatibility of extremely tough alginate/ polyacrylamide hydrogels. *Biomaterials* 34, 8042–8048 (2013).
- Yuk, H. & Zhao, X. A new 3D printing strategy by harnessing deformation, instability, and fracture of viscoelastic inks. Adv. Mater. 30, 1704028 (2018).

- 34. Jamshidi, A., Beigi, F. A. K., Kabiri, K. & Zohuriaan-Mehr, M. Optimized HPLC determination of residual monomer in hygienic SAP hydrogels. *Polym. Test.* **24**, 825–828 (2005).
- Wang, C., Yan, Q., Liu, H.-B., Zhou, X.-H. & Xiao, S.-J. Different EDC/NHS activation mechanisms between PAA and PMAA brushes and the following amidation reactions. *Langmuir* 27, 12058–12068 (2011).
- Lim, C. Y. et al. Succinimidyl ester surface chemistry: implications of the competition between aminolysis and hydrolysis on covalent protein immobilization. *Langmuir* 30, 12868–12878 (2014).

Acknowledgements We thank J. Hu and M. Guo in the Massachusetts Institute of Technology (MIT) Mechanical Engineering Department for their help with the cell viability test; T. McClure in the MIT Materials Research Laboratory for help with the FTIR measurement; and S. Ngoy in the Brigham and Women's Hospital (BHW) Rodent Cardiovascular Physiology Core for technical support with rodent surgery. This work is supported by the US National Science Foundation (NSF; grant CMMI-1661627) and Office of Naval Research (NO0014-17-1-2920). H.Y. acknowledges financial support from a Samsung Scholarship. C.E.V. acknowledges financial support from the NSF Graduate Research Fellowship Program.

**Author contributions** H.Y. conceived the idea. H.Y. developed the materials and methods for the DST. H.Y. and C.S.N. designed the in vitro and ex vivo experiments. H.Y. conducted the in vitro and ex vivo experiments. C.E.V., H.Y. and E.T.R. designed the in vivo experiments. C.E.V. and H.Y. conducted the in vivo experiments. X.M., H.Y. and X.Z. developed the quantitative model for the DST. R.F.P. conducted histological assessments. H.Y., C.E.V., R.F.P., E.T.R. and X.Z. analysed the results. H.Y. and X.Z. wrote the manuscript with input from all authors. X.Z. supervised the study.

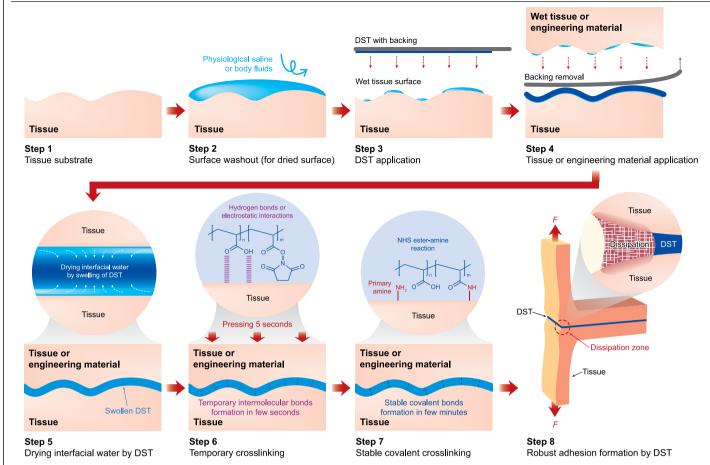
Competing interests H.Y. and X.Z. are inventors of a patent application (US patent 62/845,976) that covers the dry-crosslinking mechanism and the design of DST.

#### Additional information

**Supplementary information** is available for this paper at https://doi.org/10.1038/s41586-019-1710.5

Correspondence and requests for materials should be addressed to X.Z.

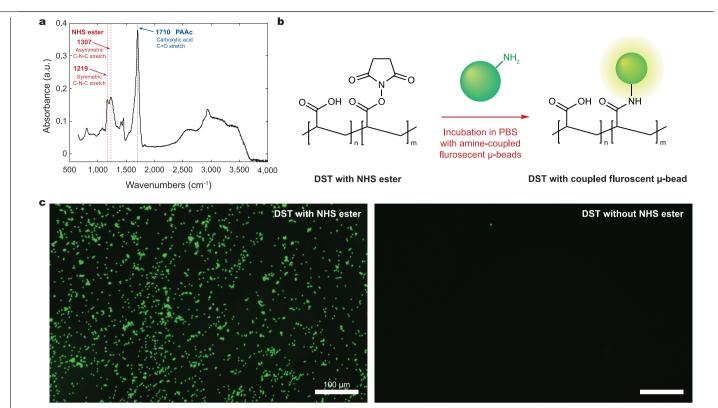
Reprints and permissions information is available at http://www.nature.com/reprints.



**Extended Data Fig. 1**| **The overall process of DST application.** The DST can be applied directly onto wet tissue surfaces after removing the backing, without any other preparation process. Upon contact with the wet surfaces, the DST dries them by quickly swelling and absorbing the interfacial water.

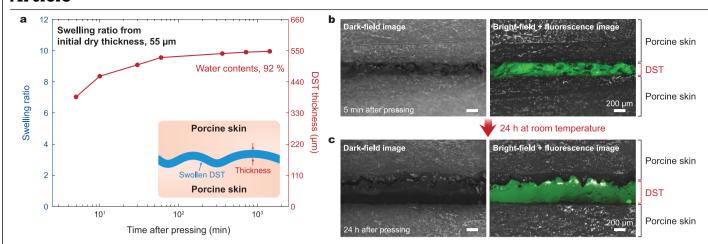
Simultaneously, the carboxylic acid groups in the DST form temporary crosslinks through intermolecular bonds with the tissue surfaces, followed by

covalent crosslinking between NHS ester groups in the DST and primary amine groups on the tissues. After adhering on tissues, the swollen DST becomes a thin layer of tough hydrogel which provides robust adhesion between the surfaces. Dissipation indicates the mechanical energy dissipation from the DST during deformation by the applied force, F.



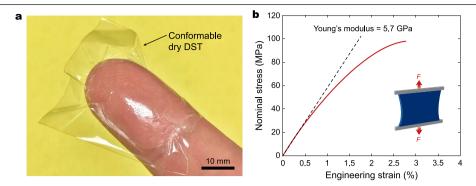
 $\label{eq:continuous} \textbf{Extended Data Fig. 2} \ | \ Presence of NHS ester and surface amine coupling of the DST. a, \ Transmission FTIR spectrum of the DST with NHS ester. The carboxylic acid C=O stretch at 1,710 cm^{-1} is associated with PAAc in the DST. The symmetric C-N-C stretch at 1,219 cm^{-1} and asymmetric C-N-C stretch at 1,307 cm^{-1} are associated with NHS ester in the DST^{35,36}. \textbf{b}, Schematic$ 

illustration of covalent crosslinking between amine-coupled fluorescent microbeads ( $\mu$ -beads) with the DST.  $\mathbf{c}$ , Fluorescence microscopy images of covalently crosslinked microbeads with DST with NHS ester (left) and DST without NHS ester (right).

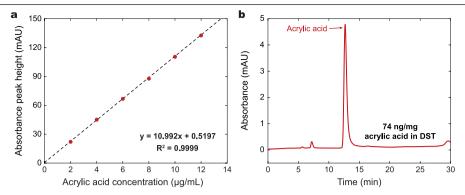


 $\label{limited} \textbf{Extended Data Fig. 3} | \textbf{Swelling of the DST. a}, \textbf{Swelling ratio} \ and \ thickness \ of the chitosan-based DST as a function of time after pressing between two wet porcine skins. \textbf{b}, Dark-field microscopic image and bright-field image overlaid with green fluorescence of porcine skins adhered by the chitosan-based DST \\$ 

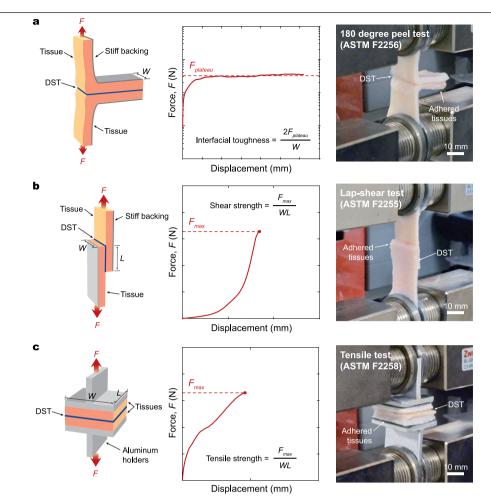
 $5\,min\,after\,application.\,c$  , Dark-field microscopic image and bright-field image overlaid with green fluorescence of porcine skins adhered by the chitosan-based DST  $24\,h\,after\,application$ . The sample was kept in a wet environment throughout the measurement.



Extended Data Fig. 4 | Properties of the dry DST. a, The DST is initially prepared as a thin dry film that can conform to tissue surfaces. b, Nominal stress versus engineering strain for the dry DST. The measured Young's modulus of the dry DST is 5.7 GPa.

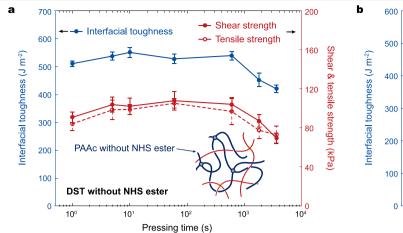


Extended Data Fig. 5 | Quantification of residual monomer in the DST by HPLC. a, Standard calibration curve of acrylic acid for HPLC. b, Results of HPLC characterization of the DST extraction solution. The DST has a very low concentration of residual acrylic acid monomers: 74 ng per 1 mg of the DST.

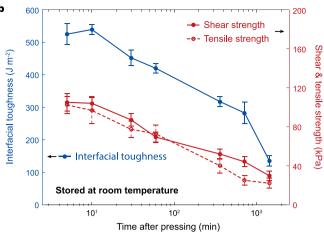


 $\label{lem:extended} \textbf{Extended Data Fig. 6} | \textbf{Setups for mechanical testing of adhesion} \\ \textbf{performance. a}, \textbf{Setup for measurement of interfacial toughness}, \textbf{based on the standard 180-degree peel test (ASTM F2256). b}, \textbf{Setup for measurement of shear strength, based on the standard lap-shear test (ASTM F2255). c}, \textbf{Setup for the standard lap-shear test} \\ \textbf{Cast MF2255} \\ \textbf{Cast MF2$ 

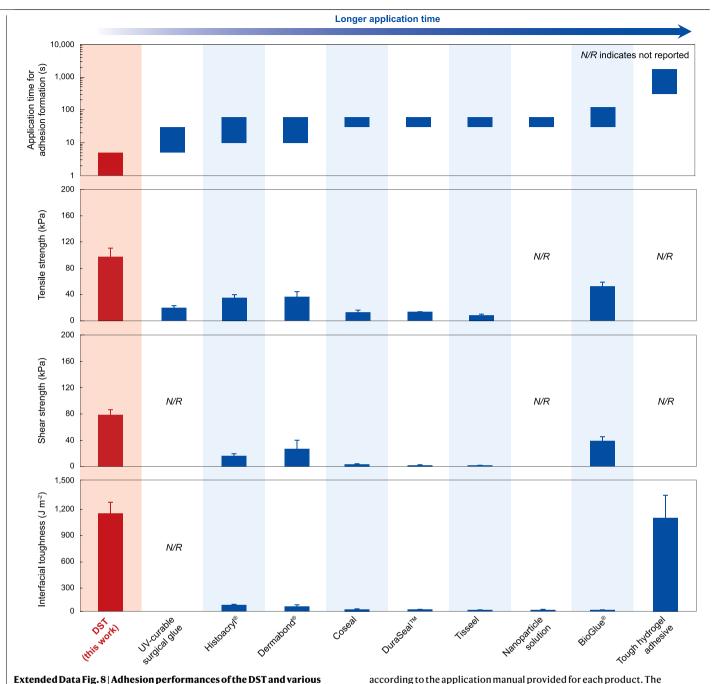
measurement of tensile strength, based on the standard tensile test (ASTM F2258). F, force;  $F_{\rm plateau}$ , plateau force in peel test;  $F_{\rm max}$ , maximum force in lap-shear and tensile test; L, length; W, width.





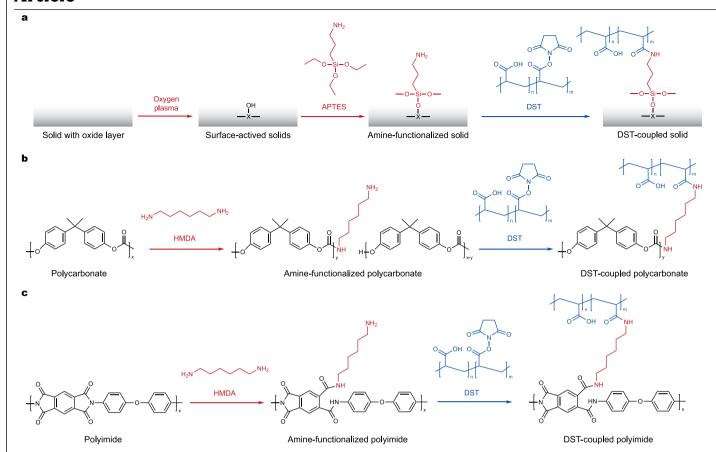


versus time after pressing for wet porcine skins adhered by the DST without NHS ester. Note that all samples for  $\bf b$  were kept in wet environments before the mechanical tests. Values represent the mean and the standard deviation (n = 3-5).



Extended Data Fig. 8 | Adhesion performances of the DST and various existing tissue adhesives. Shown are typical values for interfacial toughness, shear and tensile strength, and application time required for adhesion formation, for the DST (adhered between hydrogel and porcine skin) and various existing tissue adhesives. The interfacial toughness, shear strength and tensile strength for all commercially available adhesives were measured

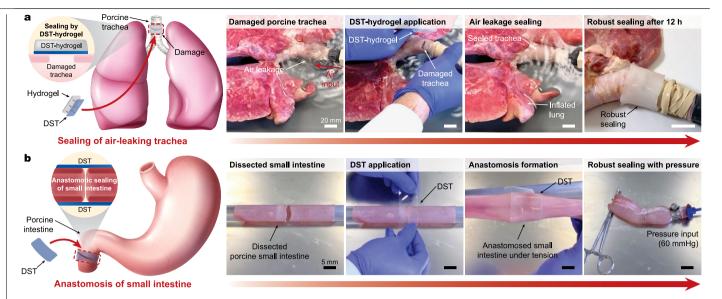
according to the application manual provided for each product. The application time for commercially available adhesives was based on the application manuals provided. The data for ultraviolet-curable surgical glue anoparticle solution and tough hydrogel adhesive were obtained from the literature. Values represent the mean and the standard deviation (n=3-5).



## $Extended\,Data\,Fig.\,9\,|\,Surface\,functionalization\,of\,engineering\,solids.$

 $\label{eq:approx} \textbf{a}, Primary amine functionalization of silicon, titanium and PDMS, and subsequent covalent coupling between the primary amine groups and the NHS ester groups in the DST. \textbf{b}, Primary amine functionalization of polycarbonate, and the primary amine functionalization of polycarbonate and the$ 

and subsequent covalent coupling between the primary amine groups and the NHS ester groups in the DST.  $\mathbf{c}$ , Primary amine functionalization of polyimide, and subsequent covalent coupling between the primary amine groups and the NHS ester groups in the DST.



**Extended Data Fig. 10 | Sealing of ex vivo porcine trachea and small intestine by the DST. a**, Sealing of an air-leaking, lacerated ex vivo porcine trachea by a hydrogel patch adhered by the DST. **b**, Anastomosis of a dissected ex vivo porcine small intestine by the DST.



Corresponding author(s):	Xuanhe Zhao
Last updated by author(s):	Aug 6, 2019

# **Reporting Summary**

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see <u>Authors & Referees</u> and the <u>Editorial Policy Checklist</u>.

Statistics				
For all statistical ar	nalyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.			
n/a Confirmed				
☐ ☐ The exact	sample size $(n)$ for each experimental group/condition, given as a discrete number and unit of measurement			
A stateme	ent on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly			
The statis Only comm	tical test(s) used AND whether they are one- or two-sided non tests should be described solely by name; describe more complex techniques in the Methods section.			
A descrip	tion of all covariates tested			
A descrip	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons			
A full desc	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)			
For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i> ) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted Give P values as exact values whenever suitable.				
For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings				
For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes				
Estimates	of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated			
·	Our web collection on <u>statistics for biologists</u> contains articles on many of the points above.			
Software an	d code			
Policy information	about <u>availability of computer code</u>			
Data collection	No software used for data collection.			
Data analysis	All data analyses have conducted by using custom codes based on MATLAB.			
For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.				
Data				
All manuscripts m - Accession code - A list of figures	about <u>availability of data</u> sust include a <u>data availability statement</u> . This statement should provide the following information, where applicable: s, unique identifiers, or web links for publicly available datasets that have associated raw data f any restrictions on data availability			
All data is available i	n the main text or the supplementary information.			
Field-spe	ecific reporting			
Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.				
☑ Life sciences       ☐ Behavioural & social sciences       ☐ Ecological, evolutionary & environmental sciences				
For a reference copy of the document with all sections, see <a href="mailto:nature.com/documents/nr-reporting-summary-flat.pdf">nature.com/documents/nr-reporting-summary-flat.pdf</a>				

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Ex vivo experiments on porcine organs and tissues were conducted to evaluate adhesion performance of the DST. The appropriate sample size Sample size (n=3-5) was used for each test.

> In vivo experiments on rat were conducted to investigate in vivo biocompatibility and biodegradability based on histological assessment after implantation. The appropriate sample size (n=3) was used to evaluate biocompatibility and biodegradability of each sample.

Data exclusions No data was excluded for ex vivo experiments.

> The following rats were also excluded: Animals that did not survive the surgery, animals that showed infection or opened the sutured incision, and animals with defective samples.

Ex vivo studies for mechanical characterization of the DST were reliably reproduced. The average and standard deviation were reported for Replication each test

In vivo studies for biocompatibility and biodegradability were reliably reproduced based on similar histological assessment for each case by the blinded pathologist. Adhesion performance was compromised when sample was defective due to expired chemical.

Randomization

No formal randomization was used but surgeries were carried out on groups, which were alternated. Each group was completed over 2-3 different surgery days.

Blinding

All histological assessments were conducted by the blinded pathologist without informing type or group of samples.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems		Me	Methods	
n/a	Involved in the study	n/a	Involved in the study	
$\boxtimes$	Antibodies	$\boxtimes$	ChIP-seq	
$\boxtimes$	Eukaryotic cell lines	$\boxtimes$	Flow cytometry	
$\boxtimes$	Palaeontology	$\boxtimes$	MRI-based neuroimaging	
	Animals and other organisms			
$\boxtimes$	Human research participants			
$\boxtimes$	Clinical data			

## Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals For ex vivo studies, porcine organs and tissues were purchased and used from Sierra Medical Inc. (Whittier, CA).

For in vivo studies, female Sprague Dawley rats, aged by weight (225-275g), were purchased from Charles River Laboratories

(Wilmington, MA).

Wild animals This study does not involve wild animals.

Field-collected samples This study does not involve field-collected samples.

Ethics oversight Both ex vivo and in vivo animal procedures were reviewed and approved by the Massachusetts Institute of Technology

Committee on Animal Care (CAC).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

# Seeing mesoatomic distortions in softmatter crystals of a double-gyroid block copolymer

https://doi.org/10.1038/s41586-019-1706-1

Received: 19 May 2019

Accepted: 23 August 2019

Published online: 28 October 2019

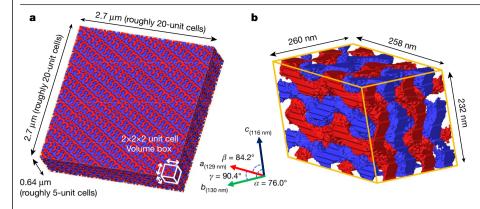
Xueyan Feng<sup>1</sup>, Christopher J. Burke<sup>2</sup>, Mujin Zhuo<sup>1</sup>, Hua Guo<sup>1</sup>, Kaiqi Yang<sup>1</sup>, Abhiram Reddy<sup>2</sup>, Ishan Prasad<sup>3</sup>. Rong-Ming Ho<sup>4</sup>. Apostolos Avgeropoulos<sup>5</sup>. Gregory M. Grason<sup>2\*</sup> & Edwin L. Thomas1\*

Supramolecular soft crystals are periodic structures that are formed by the hierarchical assembly of complex constituents, and occur in a broad variety of 'softmatter' systems<sup>1</sup>. Such soft crystals exhibit many of the basic features (such as threedimensional lattices and space groups) and properties (such as band structure and wave propagation) of their 'hard-matter' atomic solid counterparts, owing to the generic symmetry-based principles that underlie both<sup>2,3</sup>. 'Mesoatomic' building blocks of soft-matter crystals consist of groups of molecules, whose sub-unit-cell configurations couple strongly to supra-unit-scale symmetry. As yet, high-fidelity experimental techniques for characterizing the detailed local structure of soft matter and, in particular, for quantifying the effects of multiscale reconfigurability are quite limited. Here, by applying slice-and-view microscopy to reconstruct the micrometrescale domain morphology of a solution-cast block copolymer double gyroid over large specimen volumes, we unambiguously characterize its supra-unit and sub-unit cell morphology. Our multiscale analysis reveals a qualitative and underappreciated distinction between this double-gyroid soft crystal and hard crystals in terms of their structural relaxations in response to forces—namely a non-affine mode of sub-unitcell symmetry breaking that is coherently maintained over large multicell dimensions. Subject to inevitable stresses during crystal growth, the relatively soft strut lengths and diameters of the double-gyroid network can easily accommodate deformation, while the angular geometry is stiff, maintaining local correlations even under strong symmetry-breaking distortions. These features contrast sharply with the rigid lengths and bendable angles of hard crystals.

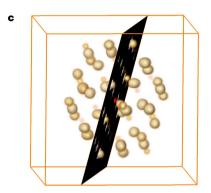
Three-dimensional (3D) tomographic imaging is the definitive experimental technique for determining the morphology of complex nanostructures. Tomography can be performed with a variety of microscopic methods, provided that there is a suitable match between the imaging resolution and the feature size. For block copolymer (BCP) structures in which periodicities are typically in the 10-100-nm regime, with domain features on the scale of 2-20 nm, electron microscopy techniques are generally required. To date, nearly all 3D tomograms of bulk-phase BCPs have been made using transmission electron microscopy (TEM) tomography<sup>4-6</sup>. Although powerful, this technique is quite limited with regard to the range of sample thicknesses, and inevitably incurs both sample deformation from microtomy and information loss associated with the restriction of tilt angles<sup>7</sup>.

Here we use the considerable advantages afforded by a wholly distinct approach to the tomography of nanostructured 3D morphologies: slice-and-view scanning electron microscopy (SVSEM; also named focused ion beam scanning electron microscopy, or FIB-SEM)8-10. Crucially, by comparison with TEM tomography, SVSEM tomography can provide a much larger reconstruction in all three spatial dimensions, facilitating 3D analysis of volumes many orders of magnitude larger than those of the typical unit cell and allowing 3D fast Fourier transform (FFT) from selected volumes within the overall reconstruction. We study a polystyrene-polydimethylsiloxane (PS-PDMS) double-gyroid BCP. The double gyroid is composed of two independent, interpenetrating enantiomorphic tubular networks of one type of block (PDMS), separated by a slab-like domain<sup>11,12</sup> (whose shape is loosely approximated by the G minimal surface<sup>13</sup>) that is constituted by the second, majority block (PS). Although a double gyroid would nominally be classified as cubic (cDG; space group of  $Ia\overline{3}d$ ) in accordance with equilibrium theories of BCPs<sup>14,15</sup>, a more critical analysis of the morphology

Department of Material Science and Nano Engineering, Rice University, Houston, TX, USA. 2Department of Polymer Science and Engineering, University of Massachusetts, Amherst, MA, USA. <sup>3</sup>Department of Chemical Engineering, University of Massachusetts, Amherst, MA, USA. <sup>4</sup>Department of Chemical Engineering, National Tsing Hua University, Hsinchu, Taiwan. <sup>5</sup>Department of Materials Science Engineering, University of Ioannina, University Campus Dourouti, Ioannina, Greece, \*e-mail; grason@mail.pse.umass.edu; elt@rice.edu



d



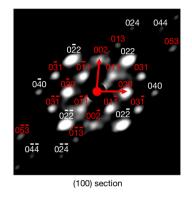


Fig. 1 | Supra-unit cell structure of the PS-PDMS double-gyroid tubular network phase. a, Real-space SVSEM reconstruction of PDMS domains (rendered red and blue), corresponding to about 2,000 unit cells of the (non-cubic) double gyroid. The white box at the bottom right highlights the size of a 2 × 2 × 2 unit-cell subvolume. b, A magnified view showing a different triclinic 2 × 2 × 2 volume cropped from within the large volume shown in panel a. The binarized (PS versus PDMS) raw SVSEM voxels are further divided into red and blue PDMS voxel networks, with the PS domains rendered transparent. The triclinic unit cell parameters  $(a, b, c, \alpha, \beta)$  and  $(\alpha, \beta)$  measured from real space are also shown. c, Rendering of the 3D FFT analysis of a region containing approximately 160 unit cells, with Bragg-like spots highlighted as quasispheriodal volumes, showing how the data intersect with a (100) plane. The central 000 peak is indicated as ared spot. See Supplementary Video 2 for an animated view of the 3D diffractogram. d, 2D logarithmic intensity plot of a (100) section from the 3D Fourier data. In addition to the allowed {220}<sub>tpg</sub>, {400}<sub>tpg</sub>, {420}<sub>rpg</sub> and {440}<sub>rpg</sub> reflections (indexed in white), there are  $\{110\}_{tDG}$ ,  $\{200\}_{tDG}$ ,  $\{310\}_{tDG}$  and  $\{530\}_{tDG}$ reflections that would be forbidden in the cubic double gyroid (indexed in red).

reveals that the slow solution-cast material organizes into distinct triclinic variants of the cubic phase.

A 3D rendering of a  $2.70 \times 2.70 \times 0.64 \,\mu\text{m}^3$  volume containing about 2,000 unit cells from within a large double-gyroid monodomain is shown in Fig. 1a. The two constituent tubular PDMS networks (shown in red and blue)—which each enclose about 20% of the total volume are defect-free, consistent with their disjointed segmentation. Prior standard small-angle X-ray scattering (SAXS) and TEM analysis indicated that this sample has a cDG structure<sup>16</sup>; however, careful analysis of the morphology throughout the 'ultra-large' volumes accessible to 3D SVSEM reconstruction reveals a decidedly non-cubic unit-cell symmetry. Figure 1b shows an experimental reconstructed 2×2×2 unit-cell volume, including triclinic unit-cell parameters. The symmetry of the particular cells in this grain deviates greatly from cubic, with the largest and smallest lattice parameters differing by 12% (for example, 130 nm versus 116 nm) and with pairs of translation vectors deviating by up to 14° from orthogonality. The synchronized slices normal to the [001] direction from the experimental reconstruction and from the corresponding deformed self-consistent field (SCF) double-gyroid model are compared in Supplementary Video 1. As shown in Extended Data Fig. 1, unit-cell parameters exhibit only small deviations throughout a given many-cubic micrometre-scale grain. Other grains exhibit distinct triclinic variants. These triclinic variants deviate from cubic symmetry by up to about 20% in both length and angle. We denote this morphology as 'variable triclinic double gyroid' (vtDG) in order to indicate unit cells that are essentially coherent within grains, but vary substantially from grain to grain. As shown in Extended Data Fig. 2, directions and magnitudes of deviations from cubic symmetry in distinct regions of the sample are uncorrelated with slicing directions, ruling out the possibility that the measured anisotropy is an artefact of SVSEM imaging or reconstruction.

Structural symmetries can also be assessed using 3D FFT of the SVSEM data. The detailed distribution of intensity from a particular (hkl) Bragg plane depends on the orientation and spacing distributions of the (hkl) planes within the volume of the sample transformed. We use selected

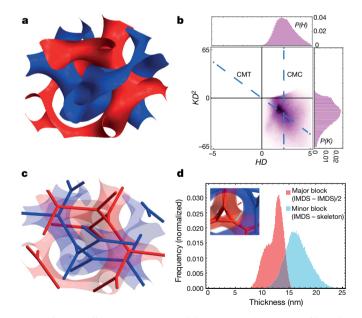


Fig. 2 | Sub-unit cell IMDS curvature and distance metrics. a, Bragg-filtered reconstruction of the IMDS for one unit cell (for a description of the IMDS isosurface, see Methods). **b**, Graph plotting the normalized Gaussian (K) and mean (H) curvatures of the IMDS and their respective probability functions (P), based on a region of 35 unit cells. The curvatures are normalized by < D > = 130 nm(the average lattice parameter cell dimension measured by SAXS). The diagonal and vertical dashed blue lines indicate the curvature distributions for a constant minimal G surface thickness (CMT) IMDS and a constant mean curvature (CMC) IMDS. c, The same unit cell as in panel a but with the IMDS made semitransparent, revealing the two skeletal graphs that are found by a thinning algorithm (see Methods). d. Distance distributions for the minority and majority  $domains. The {\it minority-block} thickness is {\it measured} from {\it the IMDS} to {\it the closest}$ distance to the skeletal graph. The majority-block thickness is measured as half the distance from a point on the red IMDS to the closest point on the blue IMDS.

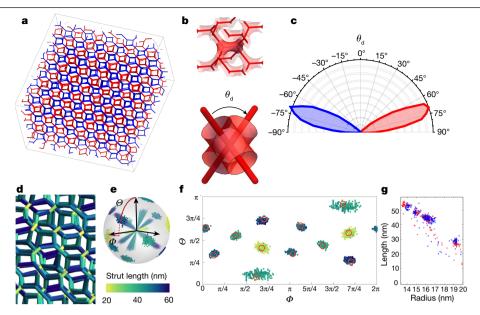


Fig. 3 | Sub-unit cell length and angular metrics. a, Region of roughly 100 unit cells, where a topological thinning of the segmented SVSEM tomogram has beenused to create the two skeletal graphs of the double-gyroid structure. The viewing direction is approximately [100]. **b**. A small region from the theoretical cubic unit cell, depicting the skeletal graph and the surrounding IMDS, highlighting an internode strut with one cubic unit cell (top) and its dihedral rotation when viewed along the strut (bottom). The solid dark red central IMDS piece contains two nodes of the graph, and by viewing along the strut connecting the nodes, one can measure the dihedral angle,  $\theta_d$ ,  $\mathbf{c}$ , Polar plot of the dihedral angle for the red ( $<\theta_d>=+70.9^\circ$ ) and blue ( $<\theta_d>=-70.8^\circ$ ) experimental networks, showing the narrow distribution of angles in each

 $network. \ \boldsymbol{d}, A portion of the skeletal graph from panel \ \boldsymbol{a}, with the struts coloured$ according to their length, showing a factor of roughly three in variability.  $\pmb{e}, Spherical \ plot \ of the \ strut \ lengths \ versus \ their \ orientation \ in \ the \ laboratory$ frame. **f**. The same data projected onto a Mercator plot, where  $\Theta$  and  $\Phi$  are the polar and azimuthal angles, respectively, of the strut orientation as shown in e. The <110> strut directions of a cDG lattice are shown as red circles. g, Inverse  $correlation\,between\,strut\,length\,and\,mean\, 'tube'\,radius\,of\,the\,IMDS\,measured$ at the midpoint along the strut (with red and blue colours indicating two distinct networks), showing the transverse contraction (dilation) upon length stretching (compression) of tubular struts.

volume diffraction (SVD)-the SEM analogue of the selected area diffraction (SAD) used in TEM analysis-in order to choose the location, shape and size of the volume to be transformed from within the larger reconstructed sample volume (see Methods). The most intense allowed cubic  $la\bar{3}d$  reflections—the {211} and {220} families—are used to define a transformation matrix that fits a triclinic lattice in order to maximize the overall intensity values at the deformed reciprocal lattice points (Fig. 1c). Measurement of the distorted reciprocal cell parameters from SVD are in perfect correspondence with the real-space measurements. Notably, inspection of the 3D SVD pattern shows intensity spots located at 'symmetry-forbidden' Bragg positions if indexed with the cubic  $Ia\overline{3}d$ space group. Forbidden reflections (see, for example, Fig. 1d) become allowed when distortions break the centring translation, screw and glide symmetries of the cubic structure. Although the experimental intensities of the most prominent of these 'forbidden' reflections are two to three orders of magnitude below those of the {211} reflections, they are 10-10<sup>5</sup> greater in magnitude than the 'symmetry-allowed' {321} and {400} reflections. The occurrence of these relatively strong 'forbidden' reflections indicates that the vtDG morphology does not simply correspond to an affine deformation of the cDG morphology, but rather to the non-affine rearrangement of the morphology at the sub-unit-cell scale. We note that distortions (attributed to solvent shrinkage forces) have resulted in the appearance of forbidden reflections in prior SAXS studies of double-gyroid structures in both bulk and thin-film BCPs 17-20.

We also analyse the sub-unit-cell morphology of the PS-PDMS double gyroid, first focusing on the shape of the intermaterial dividing surface (IMDS) and on domain thicknesses. Although the resolution of the raw SVSEM tomogram is limited by the roughly 3-nm width of image voxels (as seen in Fig. 1b), the intragrain coherence of 3D morphology over large multicell volumes enables quantitative analysis of the 'average' unit cell at higher resolution. This is accomplished by Fourier averaging of the raw greyscale SVSEM data through application of a 3D Bragg filter (see Methods)—an approach that is well established in 2D highresolution TEM<sup>21</sup>. An isosurface constructed from the Bragg-filtered SVSEM data shows the two IMDS regions, each containing PDMS and enclosing roughly 20% of the unit volume (Fig. 2a), allowing measurement of the mean (H) and Gaussian (K) curvature distributions (Fig. 2b).

Heuristically, we can compare this experimental H versus K distribution with two limiting theoretical geometries: a constant matrix thickness (CMT) surface, which is surface displaced (normally) by a constant from the G minimal surface<sup>13</sup>; or a constant mean curvature (CMC) surface<sup>22</sup>. Although a CMC shape has been suggested<sup>13,22</sup> on the grounds that it minimizes the IMDS area for a fixed volume fraction 11,23. the CMT surface minimizes the entropic penalty of variable stretching of the majority component at the expense of a slight increase in the interfacial area<sup>24</sup>. The curvature distributions for mathematical cDG surfaces of both types are shown in Extended Data Fig. 3a, b. The curvature distribution of a CMC surface is localized to a vertical band HD = 2.23, while that of a CMT surface follows Steiner's linear relationship,  $HD = -(t/2D)KD^2$  or  $HD = -0.103KD^2$  (where D is the cell repeat length and t is the constant thickness of the slab-like matrix domain)<sup>1</sup>. Relative to these reference surfaces, the experimental curvature distribution is closer to the CMC distribution. Also shown in Extended Data Fig. 3c-f are the Hand K distributions for SCF theoretical calculations of cDG as a function of segregation strength. Although the IMDS shape of these model equilibrium states is always intermediate to the CMC and CMT shapes in terms of the curvature distribution, we note that the shapes are relatively CMC-like in weakly segregated gyroids and trend towards CMT-like when interblock repulsions are increased. This observation, in combination with the CMC-like distribution measured in Fig. 2b, might suggest that the experimental IMDS shape is inherited from a state in which the PS domain vitrifies and fixes the shape of the ordering structure during solvent evaporation. It remains far from clear, however, if and how closely the shape of the partially solvated and non-equilibrium

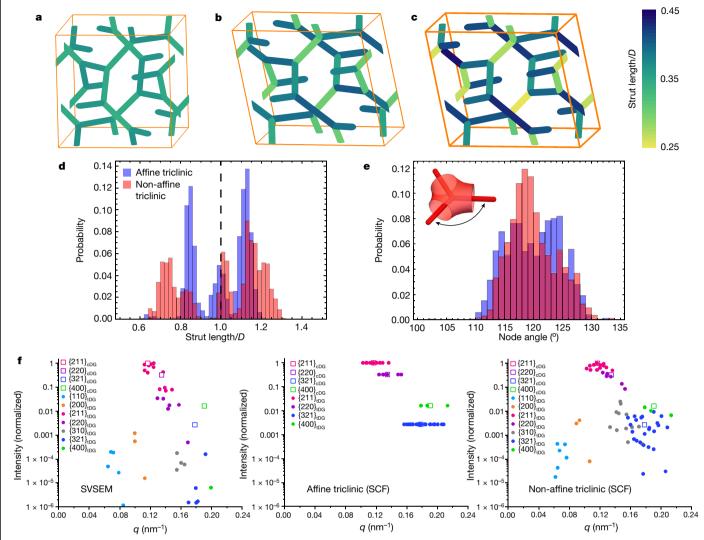


Fig. 4 | Models of sub-unit cell morphology. a-c, Orthographic views of portions of skeletal graphs for cubic (a), affine triclinic (b) and non-affine triclinic (c) models derived from equilibrium SCF calculations with cubic or triclinic symmetry. Here the colour indicates the strut length between nodes. normalized by the mean cell repeat length D.  $\mathbf{d}$ , Probability distributions for the strut length/mean cell repeat length computed for affine and non-affine triclinic

models. e, Probability distributions for internode angles for the same computed networks, with colours as in panel **d. f**, Intensity of reflections (circles) made by 3D FFT of: the SVSEM reconstruction (left; roughly 160 unit cells); the affine triclinic SCF model (centre; 64 unit cells); and the non-affine triclinic SCF model (right; 64 unit cells). For reference, the allowed peaks for cDG are plotted with open squares.

double-gyroid morphology can be modelled by an equilibrium theory for neat diblocks. We note that the IMDS shape of a BCP double gyroid has been analysed previously<sup>4-6</sup>, yet resolution limitations for the IMDS curvature of the TEM tomography measurement made identification of the surface shape signature inaccessible.

The inhomogeneous geometry of the double gyroid implies a heterogeneous distribution of domain thicknesses<sup>12,25</sup>, corresponding to variable degrees of chain extension throughout the structure. To characterize the variable thickness of the minority domains, we derive a skeletal graph from the SVSEM reconstruction; this graph consists of 1D struts threading through geometric 'centres' of tubular domains and meeting at threefold junctions (Fig. 2c)<sup>26</sup>. We define a 'minor block thickness' as the shortest distance from a point on the IMDS to the interior skeleton, while a 'major block thickness' is half of the shortest distance from a point on one IMDS to the IMDS of the opposing network (Fig. 2d, inset). Distributions of the minor (PDMS) and major (PS) block thicknesses are shown in Fig. 2d. Previous explanations for double-gyroid formation in BCP have emphasized that the constraints involved in packing polymer blocks at constant density require variation in the stretch length, most prominently derived from the greater distance from the IMDS to the threefold junction node than from the corresponding IMDS-to-sketelon distance at the mid-portion of an internode strut<sup>24,27</sup>. This notion of 'packing frustration' is consistent with the broad spread of minor-block length. However, the measured distribution of thicknesses for the major PS block also exhibits substantial spread, indicating that—contrary to the present heuristic picture<sup>24</sup> – packing of the majority block is also frustrated. Non-uniform matrix thickness thus arises as a consequence of the 'tug of war' between majority-block stretching and the countervailing forces that favour more uniform minor-block lengths, as well as from the drive towards area-minimizing IMDS shapes 23,28, consistent with the more CMC-like distribution observed in Fig. 2b.

Although the IMDS shape and domain thickness are necessarily inhomogeneous even in the ideal cDG, the internode struts in the ideal cDG are uniform in length and all orient along <110> directions. The experimental network morphology can be further analysed using the skeletal graphs (Fig. 3a). We first consider the dihedral angle (Fig. 3b). In an ideal cDG network this angle is ±70.5° (modulo 180°), where the sign characterizes the chirality of the two enantiomorphic single-gyroid

networks<sup>26</sup>. Remarkably, the experimental dihedrals for the positive and negative networks deviate little (with a root-mean-squared variance of less than 11°) from the ideal cubic geometry values (Fig. 3c). We also show (Extended Data Fig. 4) for the interstrut angles in the experimental vtDG a deviation of only about 20% from the perfect threefold coordination (120°) of cDG. This degree of local angular order in the PDMS networks contrasts starkly with the pronounced variability in the length of tubular struts as measured by skeletal edges. Strut length can vary by up to 300% (Fig. 3d). Figure 3e, fanalyses the lengths of PDMS struts according to their orientation, and indicates a strong correlation between the orientation and the 12 < 110 > directions of a cDG graph. Struts in a given orientation are relatively homogeneous in length, but show prominent length variations between distinct orientations—exceptionally large given the more modest (roughly 20%) discrepancy between triclinic and cubic cell geometry. Notably, in Fig. 3g we show concomitant contraction/ dilation of the transverse tubular radius with stretching/compression of the strut length.

To understand the origin of the anomalously large variability among PDMS strut lengths versus the relative constancy of strut angles, we consider the microdomain structures and their associated skeletal graphs derived from SCF models of the tDG. The first model, affine triclinic, is generated by affinely deforming the SCF cDG structure (Fig. 4a) into a particular tDG shape equal to that of the experimentally determined unit cell (Fig. 4b). A second model, non-affine triclinic, instead uses the same experimental triclinic cell boundary conditions to compute an equilibrium SCF double-gyroid morphology (Fig. 4c). As shown in Fig. 4d, e, the affine triclinic deformation of the double gyroid leads to spread of the strut lengths and angles by roughly 10-20%, comparable in scale to the imposed strains deforming the cubic cell to triclinic symmetry. Remarkably, if instead we consider the predicted double-gyroid morphology that equilibrates within the same triclinic cell, the network structure adopts an increased degree of length dispersity (Fig. 4d)—well beyond the nominal lengths derived from cubic to triclinic distortion—and yet a reduced degree of angle dispersity (Fig. 4e). Moreover, comparison of the spectrum of Fourier peaks from the experimental versus the SCF affine triclinic and non-affine triclinic models reveals extraordinary correspondence between the experimental structure and the non-affine triclinic model (Fig. 4f).

Taken together, these observations suggest a strong thermodynamic coupling of sub-unit-cell morphology to symmetry breaking at the supra-unit-cell scale of the double-gyroid phase. Strut lengths (and diameters) are relatively soft and thus easily accommodate deformation. presumably through relatively rapid intradomain transport of polymer chains. By contrast, the angular geometry of the gyroid network is stiff, favouring local correlations that are maintained even under strong symmetry-breaking deformations. This suggests a heuristic model for the non-affine structure of symmetry-broken soft-matter networks, consisting of periodic networks of tensed struts (so-called Steiner networks, which are 1D analogues of Plateau borders)<sup>29,30</sup> that adjust lengths locally yet maintain force-balancing angular coordination at the nodes, in order to minimize the stretching that occurs in response to imposed changes in unit-cell symmetry.

Our observations have been made possible by the accessibility of ultralarge volumes to SVSEM tomography, in combination with the Bragg averaging of selected volumes, in order to achieve enhanced resolution of sub-unit-cell features. New distance and angle metrics applied to the complex double-gyroid phase allow deeper insight into the complex energetic competition, as reflected in the distinctive structural distortions in actual samples that invariably result from solvent evaporation and grain boundary incompatibility. The symmetrybreaking distortions are predicted to have impacts on, for example, the photonic/phononic band properties of double-gyroid assemblies<sup>2,31</sup>. Our research opens up a new way of unambiguously characterizing a variety of soft-matter systems that assemble under different processing conditions into a variety of soft crystals (beyond networks), illuminating their formation mechanisms, supra-cell and sub-cell structures and structure-property relationships.

### **Online content**

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1706-1.

- Hyde, S. et al. The Language of Shape 141-197 (Elsevier Science, 1997)
- 2. Urbas, A. M., Maldovan, M., DeRege, P. & Thomas, E. L. Bicontinuous cubic block copolymer photonic crystals. Adv. Mater. 14, 1850-1853 (2002).
- Lee, J. H. et al. 25th anniversary article: ordered polymer structures for the engineering of photons and phonons, Adv. Mater. 26, 532-569 (2014).
- Laurer, J. H. et al. Microstructural analysis of a cubic bicontinuous morphology in a neat SIS triblock copolymer. Macromolecules 30, 3938-3941 (1997).
- Jinnai, H. et al. Direct measurement of interfacial curvature distributions in a bicontinuous block copolymer morphology. Phys. Rev. Lett. 84, 518-521 (2000).
- Li, Z. H. et al. Linking experiment and theory for three-dimensional networked binary metal nanoparticle-triblock terpolymer superstructures, Nat. Commun. 5, 3247 (2014).
- Ercius, P., Alaidi, O., Rames, M. J. & Ren, G. Electron tomography: a three-dimensional analytic tool for hard and soft materials research, Adv. Mater, 27, 5638-5663 (2015).
- Feng, X. Y., Guo, H. & Thomas, E. L. Topological defects in tubular network block copolymers. Polymer 168, 44-52 (2019).
- Hayworth, K. J. et al. Ultrastructurally smooth thick partitioning and volume stitching for large-scale connectomics. Nat. Methods 12, 319-322 (2015).
- 10. Narayan, K. & Subramaniam, S. Focused ion beams in biology. Nat. Methods 12, 1021-1031 (2015).
- Schick, M. Avatars of the gyroid. Physica A 251, 1-11 (1998).
- Schroeder-Turk, G. E., Fogden, A. & Hyde, S. T. Bicontinuous geometries and molecular self-assembly; comparison of local curvature and global packing variations in genusthree cubic, tetragonal and rhombohedral surfaces. Eur. Phys. J. B 54, 509-524 (2006).
- Hajduk, D. A. et al. The gyroid—a new equilibrium morphology in weakly segregated diblock copolymers. Macromolecules 27, 4063-4075 (1994).
- Matsen, M. W. The standard Gaussian model for block copolymer melts. J. Phys. Condens. Matter 14, R21-R47 (2002).
- Cochran, E. W., Garcia-Cervera, C. J. & Fredrickson, G. H. Stability of the gyroid phase in diblock copolymers at strong segregation. Macromolecules 39, 2449-2451 (2006).
- Lo. T. Y. et al. Phase transitions of polystyrene-b-poly(dimethylsiloxane) in solvents of varying selectivity, Macromolecules 46, 7513-7524 (2013).
- Toombes, G. E. S. et al. A re-evaluation of the morphology of a bicontinuous block copolymer-ceramic material, Macromolecules 40, 8974-8982 (2007).
- Faber, M., Hofman, A. H., Loos, K. & ten Brinke, G. Highly ordered structure formation in RAFT-synthesized PtBOS-b-P4VP diblock copolymers. Macromol. Rapid Commun. 37, 911-919 (2016).
- Chavis, M. A., Smilgies, D. M., Wiesner, U. B. & Ober, C. K. Widely tunable morphologies in block copolymer thin films through solvent vapor annealing using mixtures of selective solvents, Adv. Funct, Mater. 25, 3057-3065 (2015)
- Dolan, J. A. et al. Controlling self-assembly in gyroid terpolymer films by solvent vapor annealing. Small 14, 1802401 (2018).
- Henderson, R., Baldwin, J. M., Downing, K. H., Lepault, J. & Zemlin, F. Structure of purple membrane from halobacterium halobium: recording, measurement and evaluation of electron micrographs at 3.5 Å resolution. Ultramicroscopy 19, 147-178 (1986).
- Große-Brauckmann, K. Gyroids of constant mean curvature, Exp. Math. 6, 33-50 (1997).
- Thomas, E. L., Anderson, D. M., Henkee, C. S. & Hoffman, D. Periodic area-minimizing surfaces in block copolymers. Nature 334, 598-601 (1988)
- Matsen, M. W. & Bates, F. S. Origins of complex self-assembly in block copolymers. Macromolecules 29, 7641-7644 (1996).
- Schroeder, G. E., Ramsden, S. J., Christy, A. G. & Hyde, S. T. Medial surfaces of hyperbolic structures. Eur. Phys. J. B 35, 551-564 (2003).
- Prasad, I., Jinnai, H., Ho, R. M., Thomas, E. L. & Grason, G. M. Anatomy of triply-periodic network assemblies: characterizing skeletal and inter-domain surface geometry of block copolymer ayroids, Soft Matter 14, 3612-3623 (2018).
- Olmsted, P. D. & Milner, S. T. Strong segregation theory of bicontinuous phases in block copolymers. Macromolecules 31, 4011-4022 (1998).
- 28. Grason, G. M. The packing of soft materials: molecular asymmetry, geometric frustration and optimal lattices in block copolymer melts, Phys. Rep. 433, 1-64 (2006).
- Ivanov, A. O. & Tuzhilin, A. A. Minimal Networks: The Steiner Problem and Its Generalizations (CRC Press, 1994).
- Alex, J. & Grosse-Braukmann, K. Periodic Steiner networks minimizing length. Preprint at 30 http://arxiv.org/abs/1705.02471 (2017).
- Fruchart, M. et al. Soft self-assembly of Weyl materials for light and sound. Proc. Natl Acad. Sci. USA 115, E3655-E3664 (2018).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

# Methods

### Material and sample preparation

We synthesized the polystyrene-poly(dimethylsiloxane) (PS-PDMS) diblock copolymer  $^{16}$  by sequential anionic polymerization of styrene and hexamethylcyclotrisiloxane. The polymer has number average molecular weights of  $43.5\,kg\,mol^{-1}$  (for PS) and  $29.0\,kg\,mol^{-1}$  (for PDMS), with an overall composition of  $40\%\,PDMS$  (by volume) and a polydispersity index of 1.04. The sample studied was cast slowly (over the course of one week) from a 10 wt% solution (2 ml) in toluene. After drying, the sample was heated to  $60\,^{\circ}\text{C}$  for 3 days in a vacuum in order to remove any residual solvent.

On the basis of characteristic 2D TEM and radially averaged SAXS, the PS-PDMS diblock copolymer has been reported to have a double-gyroid morphology  $^{16}$ . We further characterized the small piece of sample that we used for SVSEM with the synchrotron X-ray at Sector 12-ID-B of the Advanced Photo Source in the Argonne National Laboratory. Given the SAXS pattern (Extended Data Fig. 5), the structure can indeed be nominally associated with a double-gyroid morphology, with an average cubic repeat of  $D=130\,$  nm. However, we observe a prominent low q peak associated with the  $\{110\}$  planes that is forbidden for the cubic  $la\overline{3}d$  space group. Given the size of the incident X-ray beam and the sample thickness, this SAXS pattern must come from about  $10^9\,$  unit cells. Before SVSEM imaging, the sample was attached to a  $45^\circ$  SEM stub with double-sided conductive carbon tape, and the outer surface was then coated with a 50-nm layer of platinum.

#### Slice-and-view SEM data acquisition

Extended Data Fig. 6 shows the workflow involved in SVSEM. This is a tomographic method enabled by advances in ion milling, monochromated field emission electron beams and electron imaging detectors, as well as by precise stage motion and sophisticated software routines for correlation and registration of the image stack<sup>32-34</sup>. During data collection (Extended Data Fig. 6a), ion milling is combined with electron imaging (Extended Data Fig. 7a): an incident high-energy ion beam is used to make an impact on and mill away a thin slice of the near-surface region of the sample (Extended Data SFig. 7b, i); then an electron beam is directed at the surface and a secondary-electron image is recorded (Extended Data Fig. 7b, ii). This ion-slice, electron-image sequence is repeated until a sufficient thickness of the sample has been serially imaged. The 3D tomogram is then reconstructed by vertical stacking of the aligned 2D SEM images. Three large-volume tomographic data sets from distinct double-gyroid grains are available at an online data repository (https://doi.org/10.7275/wv24-3j62).

We used a Thermo Fisher Helios NanoLab 660 SEM/FIB DualBeam system for data acquisition. A focused gallium-ion (Ga<sup>+</sup>) beam with an energy of 30 KeV and a relatively low beam current of 80 pA was used to mill the sample surface in order to minimize damage from the FIB beam<sup>35</sup>. A 1-KeV electron beam with beam current of 50 pA was used to image the sample surface with a through lens (TLD) secondary-electron detector (secondary-electron images taken with different incident energies are shown in Extended Data Fig. 8). Notably, the stronger scattering from the higher atomic number of silicon atoms in the PDMS and the resulting additional secondary-electron emission is sufficient to provide excellent intrinsic contrast between the PS and PDMS domains without staining. We used X-shaped fiducials to register the FIB and secondary-electron images during the automatic slice-and-view process, and drilled a deep hole (or holes) into the sample along the direction normal to the observing surface with FIB (at an acceleration voltage of 30 kV and beam current of 0.23 nA) for fine registration of secondary-electron images (for details, see Extended Data Fig. 9a-c). For FIB slicing, we set the slice thickness at 3 nm. Further monitoring during FIB image acquisition found the actual slice thicknesses to be  $2.96 \pm 0.01$  nm per slice (for details, see Extended Data Fig. 9d). A potential relative rotation of SEM images during SVSEM

acquisition was also monitored and excluded (Extended Data Fig. 10 and Supplementary Video 3).

### **SCFT and IMDS models**

We performed self-consistent (mean) field theory calculations of diblock copolymer melts using a polymer self-consistent field (PSCF) code (http://pscf.cems.umn.edu/) as in ref.  $^{36}$  for cubic double gyroid (cDG) and non-affine triclinic double gyroid (natDG) morphologies, with a single-chemical-parameters family: that is, a volume fraction f, product of interblock repulsion  $\chi$ , and degree of polymerization N. The initial density-field profiles for a double gyroid were based on the example files distributed with the software, and later targeted for the 40% minority (PDMS) and 60% majority (PS) volume fractions.

For the natDG we generated an initial density field that is identical to the experimentally reported lattice dimensions. We performed SCFT calculations by iteratively changing the density fields while calculating the free energy at each step and progressing towards a free-energy minimum. Then, for cDG and natDG, we further adjusted the lattice dimensions while maintaining the unit-cell length ratios in order to find a metastable density field subject to imposed symmetry constraints. For natDG, we fixed a/b = 0.97/1, a/c = 1.10/1 and angles  $\alpha$ ,  $\beta$ ,  $\gamma$  as 75.9°, 84.7°, 89.1°, which derive from the unit-cell parameters computed from the optimal (inverse) reciprocal lattice dimension from experimental SVSEM measurements, and are therefore close to those of the local volume region shown in Fig. 1b. Here we did not attempt to achieve an equivalent segregation strength  $(\chi N)$  to match PS-PDMS systems, because it is unclear which equilibrium values correspond best to the conditions in which the sample (as it is undergoing solvent evaporation) becomes ordered but remains fluid in both domains. Instead, we aimed to capture the basic mechanism of the sub-unit cell non-affine structure without imposing a close match in interblock repulsion. Hence, for fixed symmetries (cDG or natDG) we consider the cell dimension that minimizes the free energy for fixed ratios of unit-cell dimensions and angles. The affinely sheared triclinic DG (atDG) was generated from cDG, by applying a transformation matrix that consisted of lattice vectors matching the natDG at the same composition and interaction parameters. Finally, for each of the cDG, atDG and natDG, we generated image stacks of the density fields for further geometric analysis.

For the IMDS models in Extended Data Fig. 3a, b, we generated the CMC surface of a tubular double-gyroid network by using Surface Evolver<sup>37</sup> and following the procedure in ref. <sup>22</sup>. Thus we generated discrete CMC surfaces corresponding to the IMDS by minimizing the area and energy associated with a target mean curvature  $(H_0)$  for the IMDS (that is,  $(H(x) - H_0)^2 dA$ , where H(x) is the mean curvature of the surface at point x, and dA is the infinitesimal area at point x) by imposing volume constraints such that the minority component volume enclosed by the tubular networks is 40% of the cubic cell. Similarly, for the CMT IMDS, we took a discretized surface of the gyroid minimal surface and then pushed off all points by the same distance along the normal  $(\pm \hat{\bf n})$ , such that the resulting volume enclosed by surfaces on either side of the minimal surface matched the 60% volume fraction of PS in the unit cell. See the supporting software at the online data repository (https://doi. org/10.7275/wv24-3j62) for details on generating image stacks from density fields.

### **Morphological analysis**

The following subsections describe the steps (Extended Data Fig. 6b-g) that ultimately led to a detailed geometric analysis (Extended Data Fig. 6h). The custom computer codes used for these tasks are noted in each subsection, and the file 'README.txt' distributed with the supporting software outlines further instructions regarding the use of these codes with source data. All of the supporting software codes and 'README.txt' are at the online data repository (https://doi.org/10.7275/wv24-3j62).

Visualization of 3D volumes using raw secondary-electron images. See Extended Data Fig. 6b, c. Visualization was carried out with Avizo software from Thermo Fisher. The raw secondary-electron images of the PS-PDMS BCP have excellent intrinsic contrast. We further segmented the 3D tomogram by setting the initial voxel intensity threshold in order to manually define a portion of the bright, higher-intensity region (PDMS) as well as a portion of a darker, lower-intensity region (PS). We then calculated the position of the boundary between the brighter and darker regions using a gradient algorithm. The watershed operation swathen applied to fill in the two types of region. This procedure avoided the creation of any internal islands within each type of domain. After using a given threshold, we checked the volume fraction of each block against the known value from nuclear magnetic resonance (NMR: 40/60 PDMS/PS). Using these segmented images, we can reconstruct a 3D volume as coloured PDMS networks with a transparent PS matrix.

**Selected-volume diffraction of SVSEM reconstruction.** See Extended Data Fig. 6d–f. Because SVSEM captures the structure over large dimensions in every direction of the sample, it enables high-resolution analysis in reciprocal space, affording local observation of the orientation and magnitude of distinct Bragg-like intensity regions without loss of phase information. Before transforming the real-space volume data into Fourier space, we applied a Hanning window<sup>39</sup> in order to reduce artefacts in the FFT associated with discontinuities at the sample boundary. The filtered intensity of each real-space image voxel is given by:

$$I(u, v, w) = \frac{1}{8} \left[ 1 - \cos\left(\frac{2\pi(u-1)}{N_u}\right) \right] \left[ 1 - \cos\left(\frac{2\pi(v-1)}{N_v}\right) \right]$$
$$\times \left[ 1 - \cos\left(\frac{2\pi(w-1)}{N_w}\right) \right] I_0(u, v, w)$$

where  $I_0(u,v,w)$  is the original unfiltered voxel intensity;  $N_u$ ,  $N_v$  and  $N_w$  are the number of voxels in each dimension; and (u,v,w) denote the integer voxel positions in the real-space data. After applying the Hanning window, the voxel intensity data were Fourier transformed in order to obtain the reciprocal space representation, F(i,j,k).

To perform further selected-volume-diffraction (SVD) analysis, we needed an indexed reciprocal space lattice that fits with the FFT data. To achieve this, we started with a small portion of a cDG reciprocal space lattice,  $\mathbf{G}_{hkl}$ . We included the two families of non-forbidden reciprocal lattice vectors with the smallest magnitude of  $\mathbf{G}_{hkb}$  {211} and {220}. These cubic reciprocal-space vectors have a magnitude of  $|\mathbf{b}_i| = 2\pi/130 \,\mathrm{nm}^{-1}$ , where 130 nm is the average unit-cell length estimated from SAXS. In order to fit the generated lattice to the FFT data, we constructed a smooth interpolation of the FFT data, making it into a cubic spline interpolation,  $F_{\text{smooth}}(k_x, k_y, k_z)$ . The coordinates of each FFT pixel are given by  $(k_x, k_y, k_z) = (i\delta k_x, j\delta k_y, k\delta k_z)$ , where  $\delta k_x = 2\pi/\delta x$ ,  $\delta k_y = 2\pi/\delta y$ and  $\delta k_z = 2\pi/\delta z$ , and  $\delta x$ ,  $\delta y$  and  $\delta z$  are the real-space voxel dimensions. We then fit the reciprocal lattice to the FFT interpolation. In doing so, we assumed that the structure was periodic and that the real-space lattice was affinely transformed from a cubic lattice. Applying a linear transformation  $x_i \rightarrow x'_i = \Lambda_{ii} x_{ii}$ , where the nine matrix elements  $\Lambda_{ii}$  are independent, we could transform from a cubic to a triclinic unit cell. For the transformation to be linear, we required that  $k_i'x_i' = k_ix_i$ , so  $k'_i = k_i \Lambda_{ii}^{-1}$ . The goal of our fitting procedure was to find the elements of  $\Lambda^{-1}$  that transform the cubic reciprocal lattice vectors to lie on the peaks of the FFT. Specifically, we optimized the matrix elements  $\Lambda_{ii}$  in order to maximize the summed value of interpolated intensity at the deformed reciprocal lattice vectors,  $\sum_{\mathbf{k}\in\mathbf{G}_{hkl}}F_{\mathrm{smooth}}(A_{ji}^{-1}k_{j})$ , where  $\mathbf{G}_{hkl}$  are the reciprocal lattice vectors of the cubic double gyroid lattice. Having found the indexed (deformed) reciprocal lattice that fits the FFT data, we applied a targeted Bragg filter to the volume data. The Bragg filter was applied in Fourier space with a mask of Gaussian windows, while each

window was centred on the selected reciprocal-space lattice point. The Bragg mask is given by:

$$B(i,j,k) = \sum_{\mathbf{k} \in \mathbf{G}'_{hkl}} \exp \left[ \frac{(i\delta k_x - k_x)^2}{2\sigma_x^2} + \frac{(j\delta k_y - k_y)^2}{2\sigma_y^2} + \frac{(k\delta k_z - k_z)^2}{2\sigma_z^2} \right]$$

Where  $\sigma_{y}$ ,  $\sigma_{y}$  and  $\sigma_{z}$  are widths of Gaussian windows in the mask and  $\mathbf{G}'_{hbl}$ are the set of vectors in the (non-cubic) reciprocal lattice that locate each of the intensity peaks in the 3D FFT. We applied the filter pointwise to the FFT data and then applied an inverse FFT in order to obtain the filtered real-space volume data for further analysis. To carry out SVD analysis, we selected reciprocal-space lattice points on the basis of the corresponding overall intensity of each diffraction family of the FFT data and their associated cubic a value (Extended Data Fig. 11). Here we chose reciprocal-space lattice points such that their corresponding diffraction families have an overall intensity greater than 10<sup>-7</sup> (normalized by the strongest  $\{211\}_{tDG}$  family) and their associated cubic q values are smaller than 0.2 nm<sup>-1</sup> to make the Bragg filtering mask (that is, the  $\{110\}_{tDG}, \{200\}_{tDG}, \{210\}_{tDG}, \{211\}_{tDG}, \{220\}_{tDG}, \{310\}_{tDG}, \{321\}_{tDG}, \{400\}_{tDG}, \{40$ families), while the standard deviation of the Gaussian window is two pixels (that is,  $\sigma_x = 2\delta k_x$ ,  $\sigma_y = 2\delta k_y$ ,  $\sigma_z = 2\delta k_z$ ). A comparison of the 3D FFT pattern of the raw volume data and the 3D FFT pattern of the volume data after SVD treatment is shown in Extended Data Fig. 6d, e. See the supporting software 2.

Network skeletal graphs and analysis. See Extended Data Fig. 6g, h (with regard to dihedral angles, strut lengths and strut orientations). Using ImageJ (https://imagej.nih.gov/ij/), we binarized the greyscale image-stack data in order to identify the tubular networks formed by the minority domains, and separated them from the majority-block-filled matrix by using a threshold such that the volume fractions of the two binary components matched with the experimentally reported volume fractions. Note that although the same analysis could be applied to post-Bragg filtered data, the data in Fig. 3 consider a larger volume than is accessible to memory limitations of FFT filtering in Mathematica (see https://doi.org/10.7275/wv24-3j62). Hence, in order to analyse large-volume networks, we extracted the skeleton directly from the raw SVSEM data (comparative analyses of skeletons from pre- and post-filtered data in a smaller volume confirm that Bragg filtering has a negligible impact on network statistics).

We then reduced these networks into 1D skeletal graphs—that is. straight-line bonds that connect nodes which are threefold coordinated or higher; no fourfold or higher-fold nodes were identified in this way (indicating the absence of topological defects<sup>8</sup>). The initial task of reducing filtered 3D volume data into 1D lines was done using the inbuilt skeletonization feature in ImageJ. This procedure followed ref. 40, and it reduces binarized volume data into a 1D curve (also referred as a medial axis) that is a collection of voxels. To identify the skeletal graphs, we subjected the 1D curve to further refinements. We did this using a custom Mathematica code, whereby we first converted the voxel collections into a graph by taking the voxel coordinates as vertices, and then connected each voxel to its adjacent neighbours in a  $3 \times 3 \times 3$  voxel neighbourhood. For the next refinement, we fixed the vertices that lie on the boundary and iteratively removed vertices that have only one nearest neighbour which effectively removed branches of the 1D curve that did not connect to a node. Finally, we converted the remaining 1D curve into a straight line of bonds by iteratively removing vertices with two neighbours and then connecting them to one another. The end of this process usually results in having small clusters of vertices at the site of a node, which we rectified by replacing them with a single vertex, ultimately resulting in the skeletal graph with the same topology of the network that we started out with.

We then applied an optimization procedure to ensure that the skeleton lines lie along the regions of maximal density in the 3D volume data. This was achieved using an algorithm described in ref. <sup>26</sup>, which

defines an optimization functional in  $\phi = \sum_{\langle ij \rangle} \frac{1}{L_{ij}} \int_{i}^{j} ds \ \phi(\mathbf{x})$  that averages the local intensity  $\phi(\mathbf{x})$  over the skeleton bonds, where  $\langle ij \rangle$  denotes the skeleton bond connecting  $\mathbf{x}_i$  and  $\mathbf{x}_i$ , and s represents the arc length along each skeletal bond. We used a cubic spline interpolation to create the density (or intensity f) function  $\phi(\mathbf{x})$ , based on reconstructed 3D volume data from images. We maximized this function with respect to node positions x, in order to optimize the skeleton position over density  $\phi$ . We analysed the structure of the optimized skeletal graphs by calculating the skeleton dihedral angles and the length and orientation of the bonds that make up the graphs. For a given triplet of consecutive bonds, we defined the two planes and their normal as  $\hat{\mathbf{n}}_{\alpha\beta} = (\hat{\mathbf{r}}_{\alpha} \times \hat{\mathbf{r}}_{\beta})/|(\hat{\mathbf{r}}_{\alpha} \times \hat{\mathbf{r}}_{\beta})|$  and  $\hat{\mathbf{n}}_{\beta\gamma} = (\hat{\mathbf{r}}_{\beta} \times \hat{\mathbf{r}}_{\gamma})/|(\hat{\mathbf{r}}_{\beta} \times \hat{\mathbf{r}}_{\gamma})|$ , where  $\hat{\mathbf{r}}_{\alpha}$ ,  $\hat{\mathbf{r}}_{\beta}$  and  $\hat{\mathbf{r}}_{\gamma}$ are the unit vectors along the bonds. The dihedral angle is defined as the angle between these plane normals, with  $\sin\theta = (\hat{\mathbf{n}}_{\alpha\beta} \times \hat{\mathbf{n}}_{\beta\nu}) \cdot \hat{\mathbf{r}}_{\beta}$ ,  $\cos\theta = \hat{\mathbf{n}}_{\alpha\beta} \cdot \hat{\mathbf{n}}_{\beta\nu}$ . We calculated this measure for all consecutive triplets of bonds in the skeletal graph. We also calculated the bond length as  $l_{ii} = |\mathbf{r}_{ii}|$  where  $\mathbf{r}_{ii} = \mathbf{x}_i - \mathbf{x}_{ii}$  with i and j denoting the nodes (end points) of the struts, and the spherical angle coordinates  $\Theta$  and  $\Phi$  describing the length and orientation anisotropy. The node angle  $\psi = \cos^{-1}(\hat{\mathbf{r}}_{\alpha} \cdot \hat{\mathbf{r}}_{\beta})$  is computed for each of the three pairs  $(\alpha, \beta)$  of struts that meet at a single node; this is done for all nodes in the unit cell. Data from this analysis are presented as a polar histogram of dihedral angles in Fig. 3c. a histogram of node angles in Extended Data Fig. 4, Mercator plots of orientation in Fig. 3e, f, and anisotropy of strut length in Fig. 4d. See the supporting software 3, 4.

**Calculation of curvature.** See Extended Data Fig. 6g, h (with regard to IMDS curvature). We computed the mean curvature (H) and Gaussian curvature (K) of the IMDS. The IMDS is represented as a triangulated mesh, which we identified by finding a surface of the linear interpolation of density data at  $\phi_{40}$  that separated the 3D volume into three types of domain with 20%, 20% and 60% volume.

We further used two-step conditioning by first applying an edgelength regularization to the mesh, and then constraining the mesh vertices to lie on the isosurface of a third-order interpolation of the density to ensure that mesh vertices represent a surface that is at least second-order differentiable. To regularize the triangle edge lengths, we minimized a regularization functional defined as  $F_{\text{reg}} = \sum_{\langle ij \rangle} (L_{ij} - \bar{L})^2$ . We optimized this functional via a gradient-descent approach by taking the gradient with respect to the triangle vertex positions, and applied a constraint such that all resulting vertices lie on the surface by subtracting the component of gradient parallel to vertex normal. To constrain the mesh vertices, we created a third-order Hermite interpolation of density  $\phi$  and we constrained each triangle vertex to lie along the  $\phi = 0.4$ isosurfaces within this interpolation. We accomplished this by minimiz $ing(\phi - 0.4)^2$  for each vertex, again using the gradient-descent approach. We finally used the patch curvature function in the Matlab File Exchange developed by D.-J. Kroon (https://www.mathworks.com/matlabcentral/ fileexchange/32573-patch-curvature) in order to compute curvatures on the optimized triangulated mesh that represents the IMDS. This algorithm calculated the principal curvatures  $\kappa_1$  and  $\kappa_2$  associated with each triangulated vertex by fitting a paraboloid to that vertex and its nearest neighbours, with the paraboloid axis constrained along the vertex normal. From the principal curvatures, the mean curvature  $H = \frac{(\kappa_1 + \kappa_2)}{2}$  and Gaussian curvature  $K = \kappa_1 \kappa_2$  can be calculated for each vertex. Curvature-distribution data are shown in Fig. 2b and Extended Data Fig. 3. See also the supporting software 5, 6, 7.

Calculation of skeleton-IMDS and IMDS-IMDS distances and strut diameters. See Extended Data Fig. 6g, h (with regard to the

skeleton-IMDS and IMDS-IMDS distances). We used the optimized IMDS triangulated mesh and discretized skeletal graph to compute skeleton-IMDS and IMDS-IMDS distances and the effective diameters of the tubular networks. We carried out skeletal-graph discretization by choosing a discretization length  $d = \langle l_h \rangle / 100$ , where  $\langle l_h \rangle$  is the average bond length and for each bond we chose  $l_b/d$  evenly spaced points along the bond. We calculated distances between the skeleton and IMDS by finding the nearest skeleton point for each vertex on the IMDS triangulated mesh, and IMDS-IMDS distances by considering separately the two IMDS surfaces resulting from individual networks and finding the nearest vertex in one IMDS from each vertex on the other. To calculate the effective strut diameter, we found the (quasi-ellipsoidal) 1D intersection of the computed IMDS and a 2D plane that bisects a strut. An 'average tube radius' is computed by dividing the length of the 1D path (that is, the circumference) by  $2\pi$ . The skeleton-IMDS and IMDS-IMDS distances are plotted in Fig. 2d, and the effective strut diameter versus strut length in Fig. 3g. See also the supporting software 8, 9.

# **Data availability**

SVSEM and SCF modelling data are available at https://doi.org/10.7275/wv24-3j62.

### **Code availability**

Supporting software codes are available at https://doi.org/10.7275/wv24-3j62.

- Cantoni, M. & Holzer, L. Advances in 3D focused ion beam tomography. MRS Bull. 39, 354–360 (2014)
- Kotula, P. G., Rohrer, G. S. & Marsh, M. P. Focused ion beam and scanning electron microscopy for 3D materials characterization. MRS Bull. 39, 361–365 (2014).
- Wilson, J. R. et al. Three-dimensional reconstruction of a solid-oxide fuel-cell anode. Nat. Mater. 5, 541–544 (2006).
- Kim, S., Park, M. J., Balsara, N. P., Liu, G. & Minor, A. M. Minimization of focused ion beam damage in nanostructured polymer thin films. *Ultramicroscopy* 111, 191–199 (2011).
- Arora, A. et al. Broadly accessible self-consistent field theory for block polymer materials discovery. Macromolecules 49, 4675–4690 (2016).
- 37. Brakke, K. A. The surface evolver. Exp. Math. 1, 141–165 (1992).
- Beucher, S. & Meyer, F. in Mathematical Morphology in Image Processing (ed. Dougherty, E. R.) 433–481 (CRC Press, 1993).
- 39. Oppenheim, A. V. & Schafer, R. W. Discrete-Time Signal Processing (Prentice-Hall, 1999).
- Lee, T. C., Kashyap, R. L. & Chu, C. N. Building skeleton models via 3-D medial surface axis thinning algorithms. CVGIP Graph. Models Image Proc. 56, 462–478 (1994).

Acknowledgements Primary support for this research was provided through the US Department of Energy (DOE), Office of Basic Energy Sciences, Division of Materials Sciences and Engineering under award DE-SCO014599 to G.M.G. and E.L.T. Use of the Advanced Photon Source at the Argonne National Laboratory was supported by the US DOE, Office of Science, and Office of Basic Energy Sciences. A grant from the National Science Foundation to E.L.T. under award DMR 1742864 supported the development of SVSEM techniques. A grant from the Ministry of Science and Technology supported the R.-M.H. group. SCF calculations were performed on the UMass Cluster at the Massachusetts Green High Performance Computing Center. We thank B. van Leer, T. Lacon and T. Santisteban from Thermo Fisher for insights into the hardware and software underlying SVSEM tomography.

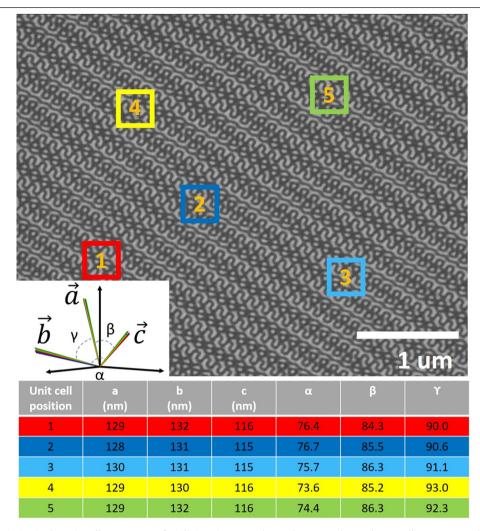
**Author contributions** The research was designed and supervised by E.L.T. and G.M.G. The BCP was synthesized by A.A., and R.-M.H. was responsible for processing samples. The SVSEM technique was developed by H.G., M. Z. and X.F., and carried out by X.F. and H.G. K.Y. provided modifications to the SVSEM software. Numerical algorithms for morphology analysis were developed by C.J.B. with I.P., and applied by C.J.B., X.F. and A.R. SCF computations were carried out and analysed by A.R. The manuscript was written by E.L.T, G.M.G., X.F., C.J.B. and A.R.

Competing interests The authors declare no competing interests.

### Additional information

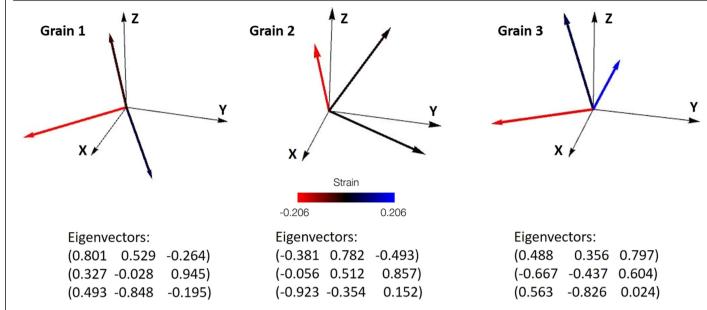
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1706-1

Correspondence and requests for materials should be addressed to G.M.G. or E.L.T. Reprints and permissions information is available at http://www.nature.com/reprints.

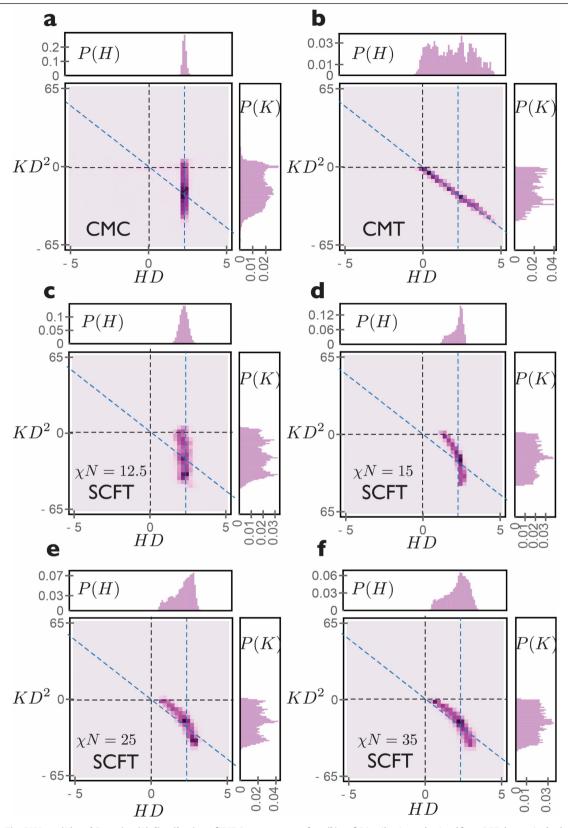


Extended Data Fig. 1 | Variation in the unit-cell parameters of triclinic unit cells within one grain. Unit-cell parameters at five different places (1-5) within one grain are measured in real space. The result indicates that the structure is

 $coherent \, but \, non-cubic, with \, unit-cell \, parameters \, exhibiting \, only \, small \, deviations \, throughout \, the \, many-cubic \, micrometre \, grain.$ 

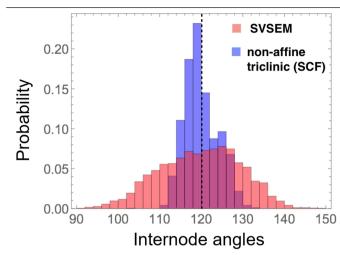


 $\textbf{Extended Data Fig. 2} | \textbf{Strain eigenvector mapping from a cDG lattice to a vtDG lattice within the slicing coordinate frame of reference.} \ \textbf{Directions and magnitudes of deviations from cubic symmetry in different grains of the sample are not correlated with the ion-milling (slicing) direction (\textit{Z}).}$ 

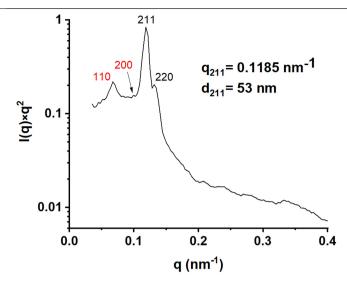


Extended Data Fig. 3 | Mean (H) and Gaussian (K) distribution of IMDS curvature in theoretical models. a, b, Distributions are shown for a constant mean curvature (CMC) surface (a) and for a constant matrix thickness (CMT)

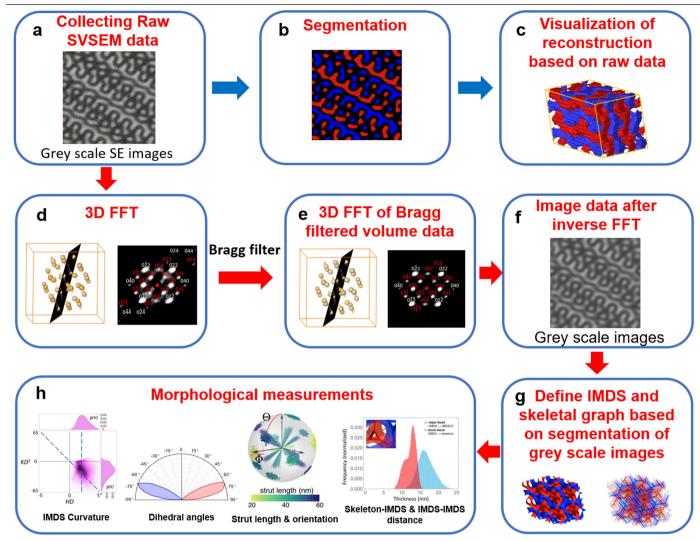
surface (**b**). **c**-**f**, Distributions obtained from SCF theoretical calculations of cDG as a function of segregation strength:  $\chi N=12.5$  (**c**),  $\chi N=15$  (**d**),  $\chi N=25$  (**e**) and  $\chi N=35$  (**f**). As in the main text, D is the cubic unit cell repeat length.



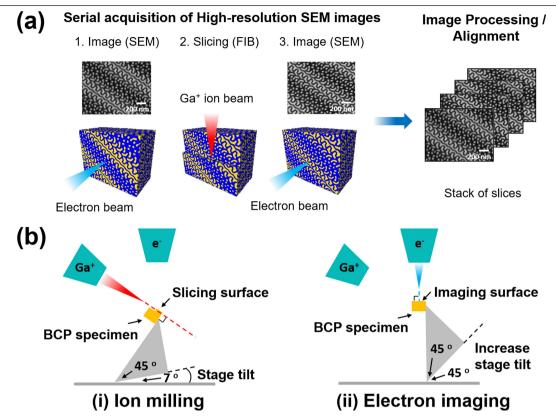
 $Extended \, Data \, Fig. \, 4 \, | \, Histogram \, showing \, the \, internode \, angles \, of \, experimental \, vtDG \, (from \, SVSEM) \, and \, non-affine \, triclinic \, SCF \, models.$ 



**Extended Data Fig. 5** | SAXS pattern from a region of the bulk polygranular **PS-PDMS sample.** The structure can be nominally associated with a double-gyroid morphology, with an average cubic lattice parameter of D=130 nm. Diffraction from the  $\{110\}_{\text{tDG}}$  and  $\{200\}_{\text{tDG}}$  families, which are forbidden for the cubic  $la\bar{3}d$  space group, are observed, indicating the non-affine deformation of the cubic double-gyroid lattice.



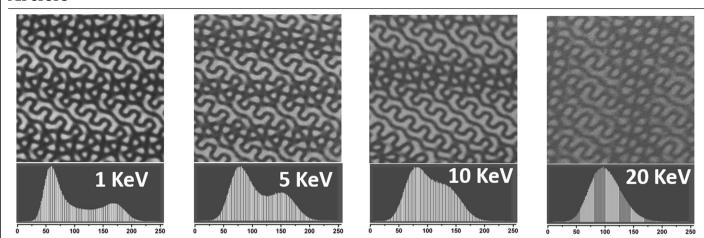
 $\textbf{Extended Data Fig. 6} | \textbf{The work flow for collection and analysis of SVSEM tomography data}. See \ \texttt{Methods for more details}.$ 



### Extended Data Fig. 7 | Acquisition and processing of SVSEM images.

 $\label{eq:approx} \textbf{a}, Illustration of the SVSEM reconstruction method. In step 1, low-energy incident electrons (1 KeV) are used to image the near-surface region of a bulk sample. In step 2, a Ga+ beam is used to slice a roughly 3-nm-thick section from the sample surface. In step 3, electrons are again used to image the ion-beam-milled sample surface. The process is repeated. With a large enough number of$ 

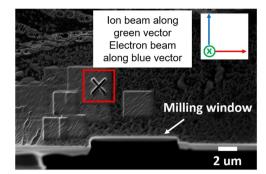
images (more than 200), the 3D morphology can be constructed via alignment of the stack of slices.  $\mathbf{b}$ , Different sample stage positions are used for undistorted imaging during slice-and-view.  $\mathbf{i}$ , For ion milling during slicing, the sample observing surface is parallel to the ion beam.  $\mathbf{ii}$ , For electron imaging, the sample observing surface is perpendicular to the electron beam.



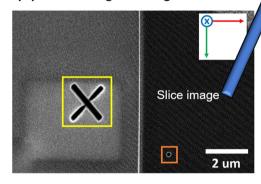
 $\label{lem:extended} \textbf{Extended Data Fig. 8} | \textbf{Secondary-electron images acquired with different electron-accelerating voltages.} \\ \textbf{Corresponding raw greyscale pixel-intensity distributions are presented blow the electron images.} \\ \textbf{With a lower accelerating} \\ \textbf{With a lower accelerating} \\ \textbf{Secondary-electron images}. \\ \textbf{With a lower accelerating} \\ \textbf{Secondary-electron images}. \\ \textbf{Secondary-electr$ 

voltage, there is a clear binary separation of the pixels into a dark peak (left) and a bright peak (right). Each image is from a freshly sliced region.

# (a) FIB Image for registration

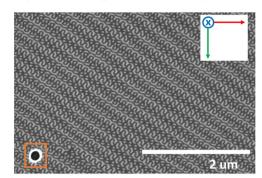


# (b) SEM Image for registration

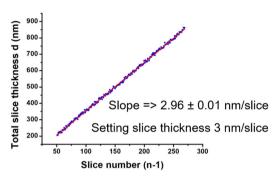


 $\label{limited} \textbf{Extended Data Fig. 9} \ | \ A lignment fiducials and monitoring of slice thickness \\ \textbf{during experiments. a}, \ An \ X-shaped fiducial (within the red square) in ion-beam \\ view is used for the registration of FIB slicing. \textbf{b}, \ An \ X-shaped fiducial (within the yellow square) in electron-beam view is used for registration of electron imaging. The round cross-section of the perpendicularly drilled hole (within the orange square) is used for the fine registration of secondary-electron images.$ 

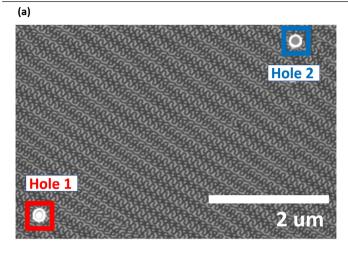
# (c) SEM Image for data acquisition

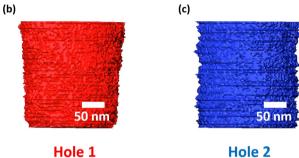


# (d) Monitoring the slice thickness

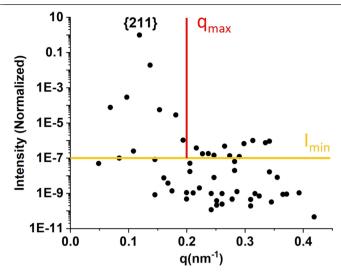


 ${f c}$ , A secondary-electron image used for data acquisition, showing an ion-milled hole (within the orange square) used for fine registration.  ${f d}$ , Monitoring of slice thickness: we measured the distance between the milling surface of the nth slice and the milling surface of the first slice (total slice thickness d) from FIB images, and then plotted d versus (n-1). This reveals a linear relationship with a slope of  $2.96 \pm 0.01$ , which is the averaged slice thickness (in nm).





Extended Data Fig. 10 | Monitoring of potential SEM image rotation during FIB-SEM image collection. a, SEM raw data image of the region of interest, with two holes drilled normal to the slice surface by the FIB. b, c, Side-view snapshots of 3D reconstructed holes 1 and 2 (the corresponding rotational videos are in Supporting Video 3). The image stack (80 slices) was aligned using hole 1. The reconstruction of hole 2 is still symmetric, indicating no image rotation.



**Extended Data Fig. 11** | **Important Fourier components of the experimental double-gyroid structure.** The overall intensity of each diffraction family was normalized by the strongest  $\{211\}_{\text{LDG}}$  family from the 3D FFT of the region containing approximately 160 unit cells (the same region as in Fig. 4f, which shows the intensity of each individual peak). The overall normalized intensity data are plotted against their associated cubic q value for those planes. For our reconstructions, we use a Bragg filter that selects peaks above an intensity threshold of  $10^{-7}$  ( $I_{\min}$ ) for associated q values smaller than  $0.2 \, \text{nm}^{-1}$ .

# California's methane super-emitters

https://doi.org/10.1038/s41586-019-1720-3

Received: 5 December 2018

Accepted: 20 August 2019

Published online: 6 November 2019

Riley M. Duren<sup>1,2\*</sup>, Andrew K. Thorpe<sup>1</sup>, Kelsey T. Foster<sup>1</sup>, Talha Rafiq<sup>3</sup>, Francesca M. Hopkins<sup>3</sup>, Vineet Yadav<sup>1</sup>, Brian D. Bue<sup>1</sup>, David R. Thompson<sup>1</sup>, Stephen Conley<sup>4</sup>, Nadia K. Colombi<sup>5</sup>, Christian Frankenberg<sup>1,6</sup>, Ian B. McCubbin<sup>1</sup>, Michael L. Eastwood<sup>1</sup>, Matthias Falk<sup>7</sup>, Jorn D. Herner<sup>7</sup>, Bart E. Croes<sup>7</sup>, Robert O. Green<sup>1</sup> & Charles E. Miller<sup>1</sup>

Methane is a powerful greenhouse gas and is targeted for emissions mitigation by the US state of California and other jurisdictions worldwide<sup>1,2</sup>. Unique opportunities for mitigation are presented by point-source emitters—surface features or infrastructure components that are typically less than 10 metres in diameter and emit plumes of highly concentrated methane<sup>3</sup>. However, data on point-source emissions are sparse and typically lack sufficient spatial and temporal resolution to guide their mitigation and to accurately assess their magnitude<sup>4</sup>. Here we survey more than 272,000 infrastructure elements in California using an airborne imaging spectrometer that can rapidly map methane plumes<sup>5-7</sup>. We conduct five campaigns over several months from 2016 to 2018, spanning the oil and gas, manure-management and wastemanagement sectors, resulting in the detection, geolocation and quantification of emissions from 564 strong methane point sources. Our remote sensing approach enables the rapid and repeated assessment of large areas at high spatial resolution for a poorly characterized population of methane emitters that often appear intermittently and stochastically. We estimate net methane point-source emissions in California to be 0.618 teragrams per year (95 per cent confidence interval 0.523-0.725), equivalent to 34–46 per cent of the state's methane inventory<sup>8</sup> for 2016. Methane 'super-emitter' activity occurs in every sector surveyed, with 10 per cent of point sources contributing roughly 60 per cent of point-source emissions—consistent with a study of the US Four Corners region that had a different sectoral mix<sup>9</sup>. The largest methane emitters in California are a subset of landfills, which exhibit persistent anomalous activity. Methane point-source emissions in California are dominated by landfills (41 per cent), followed by dairies (26 per cent) and the oil and gas sector (26 per cent). Our data have enabled the identification of the 0.2 per cent of California's infrastructure that is responsible for these emissions. Sharing these data with collaborating infrastructure operators has led to the mitigation of anomalous methane-emission activity<sup>10</sup>.

Methane (CH<sub>4</sub>) is being increasingly prioritized for near-term climate action, given its relatively short atmospheric lifetime and the potential for rapid, focused mitigation that can complement economy-wide efforts to reduce carbon dioxide emissions. In California, efforts to mitigate methane emissions are complicated by large inconsistencies between estimates of emissions derived from atmospheric measurements and from greenhouse-gas inventories: past studies using atmospheric measurements report methane emissions that are higher than those from inventories, both statewide <sup>11–13</sup> and in key regions and sectors <sup>14,15</sup>. Other studies indicate that methane emissions from the oil and gas supply chain are about 60% higher than those reported in the national greenhouse-gas inventory <sup>16</sup> and that there is a heavytail distribution of methane-emission sources in the US natural gas supply chain, where typically fewer than 20% of sources (so-called

super-emitters) contribute more than 60% of total emissions from that sector<sup>17</sup>. Scientists and policymakers have emphasized the rapid identification and mitigation of methane super-emitters, particularly those due to leaks and abnormal operating conditions<sup>18</sup>.

In addition to California, there remain large uncertainties regarding the distribution of methane emissions in other key regions and emission sectors globally<sup>19</sup>. There is a dearth of available observational studies of sectors such as livestock manure management and landfills, both of which are predicted to be larger contributors to California's methane budget than the oil and gas sector<sup>8</sup>. In addition, spatially sparse and infrequent field studies can overestimate or underestimate important methane sources that are intermittent or highly unpredictable. Finally, the relative contributions of methane point sources and area sources have not been well studied in California. We define 'point source' as a

<sup>1</sup>Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA. <sup>2</sup>University of Arizona, Tucson, AZ, USA. <sup>3</sup>University of California Riverside, Riverside, CA, USA. <sup>4</sup>Scientific Aviation, Boulder, CO, USA. <sup>5</sup>University of California Los Angeles, Los Angeles, CA, USA. <sup>6</sup>California Institute of Technology, Pasadena, CA, USA. <sup>7</sup>California Air Resources Board, Sacramento, CA, USA. \*e-mail: Riley.M.Duren@jpl.nasa.gov

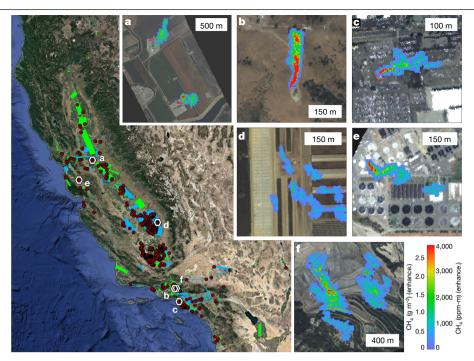


Fig. 1 | Images from our survey of methane point emissions in California. Main image, approximately 2,000 individual AVIRIS-NG flight lines from 2016 (blue) and 2017 (green) covered more than 272,000 individual facilities and infrastructure elements. Detected sources are indicated by red points, with the densest clusters seen in the San Joaquin Valley (dairies and oil fields). The inset images show examples of representative methane plumes from different sectors: a, compressor stations at a natural gas storage facility; **b**, oil well; **c**, tank of liquefied natural gas; **d**, dairy manure management;

e, wastewater-treatment plant; f, landfill. The colour scales indicate the methane concentration enhancement (the mass of methane in a plume relative to background air) in each pixel in units of parts per million-metre (ppm-m). Inset images are from AVIRIS-NG. The basemap image is from Google Earth, Lamont-Doherty Earth Observatory (LDEO)-Columbia, National Science Foundation (NSF), National Oceanic and and Atmospheric Administration (NOAA), Landsat/ Copernicus, Scripps Institution of Oceanography (SIO), US Navy, General Bathymetric Chart of the Oceans (GEBCO).

condensed surface feature or infrastructure component of less than 10 m in diameter that emits plumes of highly concentrated methane. This contrasts with an 'area source', or the combined effect of many small emitters distributed over a large area (typically 1–100 km across) that releases methane in a more diffuse fashion; area sources include anaerobic decomposition from rice cultivation and enteric fermentation from ruminant animals, both of which are better addressed with other measurement methods and are not included here.

The California Methane Survey was designed to provide the first systematic survey of methane point sources across the state, with a focus on detecting, geolocating and quantifying super-emitters. This survey fills an important gap in scale, and complements other observational systems that provide aggregate constraints on emissions from regions and area sources<sup>20–22</sup> and short-duration field campaigns that are limited to a small number of facilities 23,24. The survey was conducted with the Next Generation Airborne Visible/Infrared Imaging Spectrometer (AVIRIS-NG). AVIRIS-NG measures ground-reflected solar radiation at wavelengths from 380 nm to 2,510 nm with 5-nm spectral sampling, and has a 1.8-km field of view and 3-m pixel resolution at typical survey altitudes of 3 km (ref. 5). This class of instrument is unique in terms of its high signal-to-noise ratio, calibration accuracy and response uniformity25. The methane retrieval is based on absorption spectroscopy<sup>6,7,26</sup> and can reliably detect and quantify methane point sources with emissions typically as small as 2-10 kg CH<sub>4</sub> h<sup>-1</sup> for typical surface winds of 5 m s<sup>-1</sup>, depending on surface brightness and aircraft altitude and ground speed. See the Supplementary Information for a detailed description of datasets, estimation methods and validation.

The spatial and sectoral scope of this survey comprised key methane point-source emission sectors in California, including: oil and gas production, processing, transmission, storage and distribution; refineries; dairy manure management; landfills and composting facilities; wastewater-treatment plants; gas-fired power plants; and liquified and compressed natural gas facilities. Multiple overflights were conducted for the same infrastructure over several years to assess source persistence.

AVIRIS-NG flights for this study were conducted during five campaigns: August to November 2016. March 2017, June 2017, August to November 2017, and September to October 2018. The survey imaged approximately 59,000 km<sup>2</sup>, including revisits (Fig. 1). The survey was designed to cover at least 60% of methane point-source infrastructure in California, guided by a Geographic Information System (GIS) dataset known as Vista-CA (see Supplementary Information). Approximately 272,000 infrastructure elements were covered by the survey, most of which were observed multiple times. The survey included more than 200,000 oil and gas wells and related production infrastructure, representing a sample size more than 500 times larger than previous point-source persistence studies<sup>27</sup>.

The AVIRIS-NG flights conducted during this survey detected 1,181 individual methane plumes; for each plume we estimated the enhancement (the mass of methane in the plume relative to background air) and attributed it to a Vista-CA infrastructure element (Fig. 1). Average emission rates and 10 uncertainties were estimated for 564 distinct sources at 250 facilities, using observed methane enhancements and surface wind speed data from weather reanalysis products. The sum of our measured source emissions is 0.511 Tg CH<sub>4</sub> yr<sup>-1</sup> and we apply a nonparametric bootstrap analysis to the population of observed sources to calculate a 95% confidence interval of 0.433-0.601 Tg CH<sub>4</sub> yr<sup>-1</sup>. The population has a heavy-tail distribution, indicating that 10% of the point sources are responsible for 60% of the detected point-source emissions (Fig. 2 and Supplementary Information), spanning every sector surveyed.

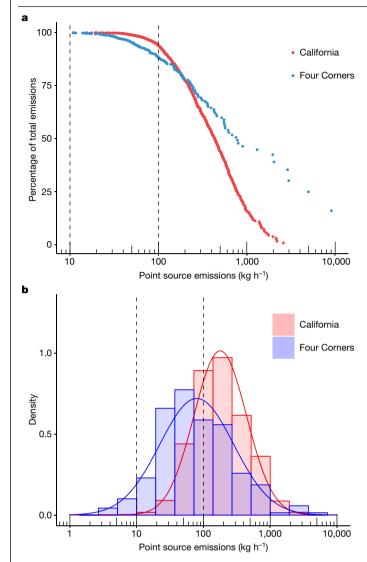


Fig. 2| The distribution of point-source emissions is consistent between two different regions. a, Data from 564 methane point sources for all sectors in California (red; this study) and from 250 coal, oil and gas sources from the Four Corners region (blue<sup>9</sup>). The numbers for California have not been adjusted for persistence here, as this was not possible for the brief Four Corners study. The heavy-tail distribution indicates that 10% of the point sources are responsible for 60% of the detected point-source emissions. b, Histogram showing the density of point-source emissions with lognormal fits. Note that the Four Corners region includes some large emitters associated with coal production that do not occur in California. The vertical dashed lines indicate typical detection limits for this class of infrared imaging spectrometer, ranging from 2–10 kg CH<sub>4</sub> h<sup>-1</sup> for the typical 3-km flight altitudes used in this study to  $100 \, \text{kg CH}_4 \, \text{h}^{-1}$  for an equivalent satellite in low Earth orbit.

The repetitive, high-spatial-resolution plume imagery enabled us to characterize point-source behaviour and controlling processes, particularly for sectors that have not been as well studied as the oil and gas production sector. Many of the sources were highly intermittent, with a median persistence of 0.20 for the entire population (mean 0.33, range 0.02–1.0). In some cases, the intermittent emissions can be explained by normal operations (for example, periodic waste flushing at large dairies). In other cases, more persistent activity is apparently due to sustained venting at a small number of anaerobic digesters at dairies and wastewater-treatment plants, or to leaking bypass valves at natural gas compressor stations. We find a similar distribution of persistence (20–35% on average) and emissions in the manure-management,

wastewater-treatment and oil and gas sectors. Solid-waste management is the largest methane point-source emission sector in California (Table 1), with persistent plumes observed at only 32 of 436 surveyed landfills and composting facilities. Our imaging of landfills identified methane plumes associated with construction, gaps in intermediate cover and leaking gas-capture wells—indicating a subpopulation of anomalous emitters (see Supplementary Information). The fact that we did not detect a larger population of smaller methane point sources across the landfill sector suggests that most of those facilities emit methane as area sources that cannot be detected with this method.

Given that we surveyed a large fraction (32–100%) of every pointsource emission sector in California, we can upscale our measurements to estimate statewide point-source emissions, resulting in a total of 0.618 (95% confidence interval 0.523-0.725) Tg CH<sub>4</sub> vr<sup>-1</sup>-equivalent to 34-46% of the California Air Resources Board (CARB) methane inventory<sup>8</sup> for 2016. We find that solid-waste management contributes 41% of observed point-source emissions, followed by 26% from manure management and 26% from oil and gas (contrasting with the 32%, 39% and 25% of total methane emissions found for these sectors in the CARB inventory8). We estimate that upstream oil and gas production contributes about 79% of the total oil and gas methane point-source emissions in California. Spatially, 85% of point-source emissions from upstream production are concentrated in the southern San Joaquin Valley (the highest oil- and associated-gas-producing region in the state), 14% in Los Angeles and Ventura counties, and 1% in the Sacramento Valley. We emphasize that the relative contribution of emission sectors probably varies in other regions around the world owing to regional differences in economic activity, age of infrastructure, and regulation. We also highlight that there are no doubt regional differences in the relative sectoral contributions of area sources (such as urban gas-distribution systems) that are beyond the scope of this study.

In addition to solid-waste management, other emission sectors may be greatly underestimated in the CARB inventory. When comparing our estimates of point-source emissions for those sectors in the CARB inventory most likely to include methane point sources, our sectoral estimates account for about 38% of the CARB inventory's emissions from the wastewater-treatment sector, about 42% of emissions from the manure-management sector, and about 366% of the CARB inventory for the energy industries sector. The latter is probably associated with most refineries and a small number of high-emitting power plants (see Supplementary Information). Large discrepancies are observed between many of the self-reported emissions from participating facilities and the AVIRIS-NG and independent airborne estimates (Fig. 3 and Supplementary Information). Moreover, our survey of point-source emissions in California and the US Environmental Protection Agency  $(EPA)'s\,Greenhouse\,Gas\,Reporting\,Program\,(GHGRP)\,for\,the\,entire\,US^{28}$ are in agreement that 99% of point-source emissions come from facilities that emit at least 25 kg h<sup>-1</sup> (see Supplementary Information). This is notable given that manure management and oil and gas production contribute more than half of the point-source emissions in our study, but are mostly not included in the GHGRP for California and are only partially represented in the total US GHGRP.

We shared preliminary findings from our surveys—including images of methane plumes—with collaborating facility operators, who provided verification with surface observations and/or explained the mechanisms underlying the observed emissions and persistence. Many of these collaborative efforts led directly to mitigation of the methane sources detected in the survey. For example, we discovered four cases of leaking natural gas distribution lines and one leaking liquified natural gas storage tank (Fig. 1), which the operators confirmed, repaired, and requested verification of repair by follow-up AVIRIS-NG flights 10.

The prevalence of methane super-emitter activity in multiple sectors in California suggests substantial potential for mitigation. We have found that 30 facilities could be responsible for around 20% of the 2016 CARB methane inventory, including many that exhibit large

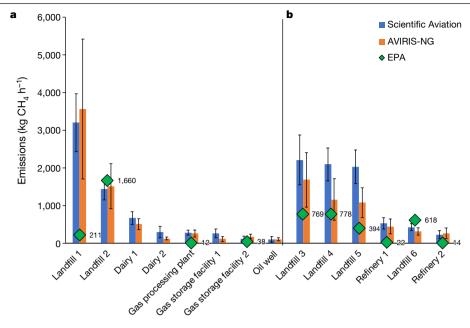
Table 1 | Point-source emissions by sector

IPCC source category	Vista-CA infrastructure element	Number of Vista-CA infrastructure elements	Number of surveyed elements	Percentage surveyed	Sectoral scalar		Measured emissions (Tg CH <sub>4</sub> yr <sup>-1</sup> )	State total emissions (Tg CH <sub>4</sub> yr <sup>-1</sup> )	State total 95% confidence intervals (Tg CH <sub>4</sub> yr <sup>-1</sup> )	Percentage of total emissions
1A1 Energy industries	Gas-fired power plants	435	238	55	1.83	7	0.007	0.013	0.007, 0.021	2.1
	Refineries	26	26	100	1.00	37	0.015	0.015	0.008, 0.023	2.4
	Subtotals	461	264	57	1.27	44	0.022	0.028	0.015, 0.044	4.6
1B2 Oil and natural gas	CNG/LNG fuelling stations	208	132	63	1.58	6	0.002	0.003	0.003, 0.004	0.5
	Natural gas stations (non-storage compressor, metering, etc)	1,131	538	48	2.10	5	0.005	0.010	0.009, 0.012	1.6
	Natural gas pipeline (transmission, distribution)	216,774	68,548	32	3.16	5	0.004	0.012	0.010, 0.014	1.9
	Natural gas processing plants	26	23	88	1.13	5	0.004	0.004	0.004, 0.005	0.7
	Natural gas storage fields	12	12	100	1.00	11	0.009	0.009	0.008, 0.010	1.4
	Oil and gas: wells	225,766	198,231	88	1.14	107	0.048	0.054	0.046, 0.063	8.8
	Oil and gas: other production equipment	3,356	2,872	86	1.00	120	0.066	0.066	0.056, 0.076	10.7
	Subtotals	447,273	270,356	60	1.16	259	0.137	0.158	0.135, 0.184	25.6
3A2 Manure management	Dairy confined animal feeding operations	620	443	71	1.40	215	0.115	0.161	0.137, 0.187	26.1
4A1 Managed waste disposal	Landfills and composting facilities	1,146	436	38	1.11	32	0.229	0.255	0.175, 0.345	41.3
4D1, 4D2 Wastewater treatment and discharge	Domestic and industrial wastewater treatment	148	57	39	2.60	12	0.004	0.012	0.005, 0.020	1.9
	Industrial wastewater treatment: beef processing	NA	NA	NA	1.00	2	0.004	0.004	0.004, 0.005	0.6
	Totals	449,648	271,556	60	1.21	564	0.511	0.618	0.523, 0.725	100.0

The table summarizes the persistence (frequency)-adjusted point-source emissions found in this study according to sectors identified by the Intergovernmental Panel on Climate Change (IPCC), as well as estimated total emissions derived with population scalars. Most of the scalars are simply the ratio of the number of infrastructure elements identified by Vista-CA to the number of surveyed elements, with three exceptions (oil and gas; other production equipment; landfills and composting facilities; and industrial wastewater treatment), for which we further constrain or eliminate scaling. See Supplementary Information section 2 for details.

discrepancies between reported and measured emissions (see Fig. 3) and Supplementary). Our survey in California and a previous study of the Four Corners region in the US9 exhibit consistent heavy-tail distributions of methane point-source emissions (Fig. 2) despite the different sectoral mixes for the two regions (the Four Corners emissions are associated primarily with oil, gas and coal production<sup>9</sup>). If similar distributions of methane point-source emissions occur in other key regions around the world, this could translate to as much as 8-11% of global greenhouse-gas forcing, assuming a 100-year warming potential of 32 and 350 Tg CH<sub>4</sub> yr<sup>-1</sup> of total anthropogenic methane emissions for 2016 (refs. 19,29). Testing this hypothesis would require additional aircraft surveys and satellite observations that can provide the necessary combination of high spatial resolution, sensitivity and wide area coverage for other key regions globally. Those broader studies would also improve our understanding of waste and manure-management emissions, which, as in California, might dominate the emission budgets of other regions19.

Detection limits for methane point sources could be relaxed by a factor of ten compared with the survey described here and still identify 90% of super-emitters if applied frequently over large areas that have emission distributions similar to those of California (Fig. 2). Because detection scales linearly with spatial resolution<sup>30</sup>, mature technologies such as that used here could be deployed for more efficient point-source monitoring across larger regions on high-altitude aircraft and satellites. Our highperformance infrared imaging spectroscopy would translate to a robust  $detection \, limit \, of \, 100 \, kg \, CH_4 \, h^{\text{--}1} for \, a \, satellite \, in \, low \, Earth \, orbit, \, depend$ ing on spatial resolution (assuming a wind speed of  $5\,\mathrm{m\,s^{-1}}$ ). Widespread and sustained deployment of point-source remote sensing methods such as ours, when combined with near-continuous regional monitoring of distributed area sources by surface observations and other satellites, could greatly advance scientific understanding of methane budgets and efforts to manage them. Complete closure of the methane budget and effective mitigation will no doubt require a multi-tiered observational strategy, in which the methods demonstrated here could play a key part.



**Fig. 3** | Independent airborne measurements of emissions from representative facilities on the basis of simultaneous flights or several visits. a, Simultaneous flights; **b**, average emissions from multiple non-simultaneous flights over several months. Orange bars show AVIRIS-NG estimates of point-source emissions, and blue bars show estimates by Scientific Aviation (Boulder, CO, USA) of facility net emissions<sup>31</sup>. Error bars indicate one

standard deviation. AVIRIS-NG estimates are lower than Scientific Aviation estimates for facilities that have some non-point-source activity. The 14 estimates here correlate with an  $R^2$  of 0.86 (see Supplementary Information). The  $R^2$  for the eight facilities in a is 0.99. The estimated total emissions here are 11,228  $\pm$  4,981 kg h<sup>-1</sup> (AVIRIS-NG) and 13,900  $\pm$  3,593 kg h<sup>-1</sup> (Scientific Aviation). Green diamonds indicate available self-reported emissions<sup>28</sup>.

### Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1720-3.

- California Senate Bill 1383. Short-Lived Climate Pollutants https://legiscan.com/CA/bill/ SB1383/2015 (2016).
- 2. Global Methane Initiative. https://www.globalmethane.org (2019).
- Zavala-Araiza, D. et al. Super-emitters in natural gas infrastructure are caused by abnormal process condition. Nat. Commun. 8, 14012 (2017).
- National Academies of Sciences, Engineering, and Medicine. Improving Characterization of Anthropogenic Methane Emissions in the United States (National Academies Press, 2018).
- Hamlin, L. et al. Imaging spectrometer science measurements for terrestrial ecology: AVIRIS and new developments. In *IEEE Aerospace Conf. Proc.* https://ieeexplore.ieee.org/document/5747395 (2011).
- Thorpe, A. K. et al. Airborne DOAS retrievals of methane, carbon dioxide, and water vapor concentrations at high spatial resolution: application to AVIRIS-NG. Remote Sens. Environ. 179, 104–115 (2016).
- Thompson, D. R. et al. Real-time remote detection and measurement for airborne imaging spectroscopy: a case study with methane. Atmos. Meas. Tech. 8, 4383–4397 (2015).
- California Greenhouse Gas Emission Inventory. Methane emissions for 2016. California Air Resources Board https://ww3.arb.ca.gov/cc/inventory/data/tables/ghg\_inventory\_by\_ipcc\_all\_00-17.xlsx (2018).
- Frankenberg, C. et al. Airborne methane remote measurements reveal heavy-tail flux distribution in Four Corners region. Proc. Natl Acad. Sci. USA 113. 9734–9739 (2016).
- Photojournal. NASA instrument detects methane gas leak. Jet Propulsion Laboratory/ California Institute of Technology https://photojournal.jpl.nasa.gov/catalog/PIA22467 (2018).
- Wecht, K. J. et al. Spatially resolving methane emissions in California: constraints from the CalNex aircraft campaign and from present (GOSAT, TES) and future (TROPOMI, geostationary) satellite observations. Atmos. Chem. Phys. 14, 8173–8184 (2014).
- Turner, A. J. et al. Estimating global and North American methane emissions with high spatial resolution using GOSAT satellite data. Atmos. Chem. Phys. 15, 7049–7069 (2015).
- Jeong, S. et al. A multitower measurement network estimate of California's methane emissions. J. Geophys. Res. Atmos. 118, 11339–11351 (2013).
- Wong, C. K. et al. Monthly trends of methane emissions in Los Angeles from 2011 to 2015 inferred by CLARS-FTS observations. Atmos. Chem. Phys. 16, 13121–13130 (2016).

- Jeong, S., Millstein, D. & Fischer, M. L. Spatially explicit methane emissions from petroleum production and the natural gas system in California. Environ. Sci. Technol. 48, 5982–5990 (2014).
- Alvarez, R. et al. Reconciling divergent estimates of oil and gas methane emissions. Proc. Natl Acad. Sci. USA 112, 15597–15602 (2015).
- Brandt, A. et al. Methane leaks from North American natural gas systems. Science 343, 733–735 (2014).
- California Assembly Bill 1496. Methane emissions. https://legiscan.com/CA/bill/ AB1496/2015 (2016).
- Saunois, M. et al. The global methane budget 2000–2012. Earth Syst. Sci. Data 8, 697–751 (2016).
- Jeong, S. et al. Estimating methane emissions from biological and fossil-fuel sources in the San Francisco Bay Area. Geophys. Res. Lett. 44, 486–495 (2016).
- Verhulst, K. R. et al. Carbon dioxide and methane measurements from the Los Angeles Megacity Carbon Project. Part 1: calibration, urban enhancements, and uncertainty estimates. Atmos. Chem. Phys. 17, 8313–8341 (2017).
- Yadav, V. et al. Spatio-temporally resolved methane fluxes from the Los Angeles Megacity. J. Geophys. Res. Atmos. 124, 5131–5148 (2019).
- Conley, S. et al. Methane emissions from the 2015 Aliso Canyon blowout in Los Angeles, CA. Science 351, 1317–1320 (2016).
- Krautwurst, S. et al. Methane emissions from a Californian landfill, determined from airborne remote sensing and in situ measurements. Atmos. Meas. Tech. 10, 3429–3452 (2017).
- Mouroulis, P. & Green, R. O. Review of high fidelity imaging spectrometer design for remote sensing. Opt. Eng. 57, 040901 (2018).
- 26. Thompson, D. R. et al. Space-based remote imaging spectroscopy of the Aliso Canyon  ${\rm CH_4}$  superemitter. Geophys. Res. Lett. 43, 6571–6578 (2016).
- Englander, J. G., Brandt, A. R., Conley, S., Lyon, D. R. & Jackson, R. B. Aerial interyear comparison and quantification of methane emissions persistence in the Bakken Formation of North Dakota, USA. *Environ. Sci. Technol.* 52, 8947–8953 (2018).
- Greenhouse Gas Reporting Program (GRP). United States Environmental Protection Agency https://www.epa.gov/ghgreporting/ghg-reporting-program-data-sets (2018).
- Etminan, M., Myhre, G., Highwood, E. J. & Shine, K. P. Radiative forcing of carbon dioxide, methane, and nitrous oxide: a significant revision of the methane radiative forcing. Geophys. Res. Lett. 43, 12614–12623 (2016).
- Jacob, D. J. et al. Satellite observations of atmospheric methane and their value for quantifying methane emissions. Atmos. Chem. Phys. 16, 14371–14396 (2016).
- Methane Hotspots Research (AB1496). Airborne Facility-Level Methane Emissions Study California Air Resources Board https://ww2.arb.ca.gov/our-work/programs/methane/ ab1496-research (2018).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

# **Data availability**

Radiance and reflectance products calibrated by AVIRIS-NG can be ordered from the AVIRIS-NG data portal at https://avirisng.jpl.nasa.gov/alt\_locator/. Retrieved methane images from flight lines in this study are available for download at https://doi.org/10.3334/ORNLDAAC/1727. Vista-CA infrastructure spatial layers are available for download at https://doi.org/10.3334/ORNLDAAC/1726. Images of methane plumes, Vista-CA layers and regional-scale methane-emission products for California can be viewed at https://methane.jpl.nasa.gov/. Tables of methane plume and source characteristics are provided in the Supplementary Information.

# **Code availability**

The custom computer code or algorithms used to generate the results in this study can be made available to researchers upon request.

Acknowledgements We thank the AVIRIS-NG team and Dynamic Aviation for their efforts in conducting the multiple airborne campaigns involved in this study, and our former colleague at the Jet Propulsion Laboratory (JPL), A. Aubrey, for early support in project planning. We acknowledge G. Franco (California Energy Commission, CEC) and E. Tseng (University of California Los Angeles) for comments on the paper. We appreciate the many discussions and input to flight planning and analysis from our colleagues at the California Air Resources Board

(CARB), the Bay Area Air Quality Management District, the South Coast Air Quality Management District, the CEC, Southern California Gas Company, Sunshine Canyon Landfill Local Enforcement Agency, and the Milk Producer's Council. We thank our colleagues at the Pacific Gas and Electric Company for their support for natural gas control release tests. We thank NASA's Earth Science Division, particularly J. Kaye, for continued support of AVIRIS-NG methane science. Additional funding for data collection and analysis was provided to JPL by CARB under ARB-NASA Agreement 15RD028 Space Act Agreement 82-19863 and the CEC under CEC-500-15-004. The data from follow-up and contemporaneous Scientific Aviation flights used in this study were funded by CARB. Analysis of this work was also supported in part by NASA's Carbon Monitoring System (CMS) Prototype Methane Monitoring System for California and the Advancing Collaborative Connections for Earth System Science (ACCESS) Methane Source Finder project. A portion of this research was carried out at the JPL, California Institute of Technology, under contract with NASA (NNN12AA01C). The authors are responsible for the content of the paper and the findings do not represent the views of the funding agencies.

**Author contributions** R.M.D., A.K.T., F.M.H. and C.E.M conceived the study. R.M.D., A.K.T., F.M.H., T.R., I.B.M., M.L.E. and S.C. conducted flight planning. Each author contributed to the collection, analysis or assessment of one or more datasets necessary to perform this study. R.M.D., A.K.T., K.T.F., F.M.H. and T.R performed the analysis with contributions from B.D.B., D.R.T., C.F., N.K.C., M.F., J.D.H, B.E.C., R.O.G. and V.Y. R.M.D., A.K.T., K.T.F. and F.M.H. wrote the manuscript with input from all authors.

Competing interests The authors declare no competing interests.

#### Additional information

**Supplementary information** is available for this paper at https://doi.org/10.1038/s41586-019-1720-3.

Correspondence and requests for materials should be addressed to R.M.D.

Reprints and permissions information is available at http://www.nature.com/reprints.

# Human origins in a southern African palaeo-wetland and first migrations

https://doi.org/10.1038/s41586-019-1714-1

Received: 30 October 2018

Accepted: 24 September 2019

Published online: 28 October 2019

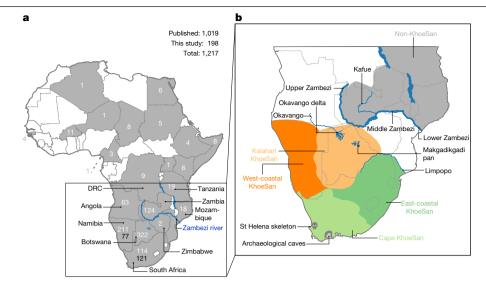
Eva K. F. Chan<sup>1,2</sup>, Axel Timmermann<sup>3,4\*</sup>, Benedetta F. Baldi<sup>1</sup>, Andy E. Moore<sup>5</sup>, Ruth J. Lyons<sup>1</sup>, Sun-Seon Lee<sup>3,4</sup>, Anton M. F. Kalsbeek<sup>1</sup>, Desiree C. Petersen<sup>1,11</sup>, Hannes Rautenbach<sup>6,7,12</sup>, Hagen E. A. Förtsch<sup>8</sup>, M. S. Riana Bornman<sup>7</sup> & Vanessa M. Hayes<sup>1,2,7,9,10</sup>\*

Anatomically modern humans originated in Africa around 200 thousand years ago (ka)<sup>1-4</sup>. Although some of the oldest skeletal remains suggest an eastern African origin<sup>2</sup>, southern Africa is home to contemporary populations that represent the earliest branch of human genetic phylogeny<sup>5,6</sup>. Here we generate, to our knowledge, the largest resource for the poorly represented and deepest-rooting maternal LO mitochondrial DNA branch (198 new mitogenomes for a total of 1,217 mitogenomes) from contemporary southern Africans and show the geographical isolation of LOd1'2, LOk and LOg KhoeSan descendants south of the Zambezi river in Africa. By establishing mitogenomic timelines, frequencies and dispersals, we show that the LO lineage emerged within the residual Makgadikgadi-Okavango palaeo-wetland of southern Africa<sup>7</sup>, approximately 200 ka (95% confidence interval, 240–165 ka). Genetic divergence points to a sustained 70,000-year-long existence of the L0 lineage before an out-of-homeland northeast-southwest dispersal between 130 and 110 ka. Palaeoclimate proxy and model data suggest that increased humidity opened green corridors, first to the northeast then to the southwest. Subsequent drying of the homeland corresponds to a sustained effective population size (LOk), whereas wet-dry cycles and probable adaptation to marine foraging allowed the southwestern migrants to achieve population growth (LOd1'2), as supported by extensive south-coastal archaeological evidence<sup>8-10</sup>. Taken together, we propose a southern African origin of anatomically modern humans with sustained homeland occupation before the first migrations of people that appear to have been driven by regional climate changes.

Southern Africa has long been considered to be one of the regions in which anatomically modern humans (AMHs) originated. Home to contemporary populations who represent the earliest human lineages, evolutionary time estimates have largely been based on mitochondrial DNA (mitogenomes)<sup>1,6</sup>. The maternal human phylogenetic tree consists of two major branches, the extensive L1'6—which includes the out-of-Africa ancestral L3 sub-branch (or haplogroup)—and the rare deep-rooting LO. The LO lineage is predominated by southern African haplogroups: LOd, LOk and the recently described LOg<sup>6</sup>. By contrast, the rare LOf and common LOa lineages are dispersed throughout sub-Saharan Africa<sup>1,3,6</sup>. Through LO pre-screening, we identified 198 southern Africans with poorly represented haplogroups for whom the mitogenome was sequenced (Supplementary Table 1), allowing for a combined analysis of 1,217 mitogenomes (Fig. 1a and Extended Data Table 1).

We ethno-linguistically classified study participants as KhoeSansouthern African populations who traditionally practiced foraging and spoke languages containing 'click' consonants—or non-KhoeSan individuals. Non-KhoeSan who have KhoeSan-derived LO mitogenomes are referred to in this study as KhoeSan ancestral, with further geographical classification (Fig. 1b and Extended Data Table 2; terminology pertaining to southern African KhoeSan populations is complex and contentious, see Methods for further discussion). Contemporary KhoeSan include Kalahari KhoeSan (Kx'a, Tuu and central Khoe-Kwadi speakers) and west-coastal KhoeSan (Khoe-Kwadi Nama speakers)<sup>11</sup>. Peoples who speak Southern Bantu languages, who migrated down the east coast of Africa around 1,500 years ago, may have acquired an east-coastal KhoeSan heritage<sup>12</sup>. The arrival of European colonists to the Cape in mid-1600s gave rise to the South African Coloured and Namibian Baster populations (of Eurasian and indigenous descent), who acquired a Cape KhoeSan heritage<sup>13</sup>. Excluding the east African Sandawe and Hadza (whose languages also contain click consonants),  $in digenous\,Khoe San\,populations\,appear\,to\,be\,absent\,nor the ast\,of\,the$  $Zambeziriver, supported by the lack of skeletal remains \, representing \, the \, in the contract of the lack of skeletal remains \, representing the \, in the lack of skeletal remains \, representing \, representing the \, in the lack of skeletal remains \, representing the \, in the lack of skeletal remains \, representing the \, in the lack of skeletal remains \, representing the \, in the lack of skeletal remains \, representing the \, in the lack of skeletal remains \, representing the \, in the lack of skeletal remains \, representing the \, in the lack of skeletal remains \, representing the \, in the \, in the lack of skeletal remains \, representing \, representing the \, in the \,$ KhoeSan-like hunter-forager morphology<sup>14</sup>. We classified the 198 new

Genomics and Epigenetics Division, Garvan Institute of Medical Research, Sydney, New South Wales, Australia. 2St Vincent's Clinical School, University of New South Wales, Sydney, New South Wales, Australia. 3 Center for Climate Physics, Institute for Basic Science, Busan, South Korea. 4 Pusan National University, Busan, South Korea. 5 Department of Geology, Rhodes University, Grahamstown, South Africa. 6Climate Change and Variability, South African Weather Service, Pretoria, South Africa. 7School of Health Systems and Public Health, University of Pretoria, Pretoria, South Africa, 8Windhoek Central Hospital, Windhoek, Namibia, 9Faculty of Health Sciences, University of Limpopo, Soyenga, South Africa, 10Central Clinical School, University of Sydney, Sydney, New South Wales, Australia. 11 Present address: The Centre for Proteomic and Genomic Research, Cape Town, South Africa. 12 Present address: Akademia, Johannesburg, South Africa, \*e-mail: timmermann@pusan.ac,kr; v.haves@garvan.org.au



**Fig. 1**| **Geographical distribution of 1,217 L0 mitogenomes. a**, Countries within (n=1,139) or outside (n=78) Africa from which L0 mitogenomes were sourced, including 198 new L0 mitogenomes (black numbers on the map). DRC, Democratic Republic of the Congo. **b**, Present-day southern Africa showing the geographical distribution of KhoeSan population identifiers defined as

 $Khoe San \, (orange), Kalahari \, and \, west-coastal; Khoe San \, ancestral \, (green), Cape \, or \, east-coastal. \, The \, Zambezi \, river \, provides \, a \, geographical \, division \, between \, the \, Khoe San \, and \, mostly \, non-Khoe San \, population \, identifiers. \, Maps \, were \, generated \, in \, the \, R \, package \, 'maps' \, v. \, 3. \, 3.0^{37}.$ 

mitogenomes as Kalahari (n=18), west-coastal (n=21), Cape (n=109) and east-coastal (n=29) KhoeSan, or non-KhoeSan (Bantu, n=19), although two mitogenomes were classed as unknown. Using these identifiers, we provide a best-fit classification for all 1,217 LO mitogenomes (Supplementary Table 2).

Phylogenetic analysis confirms the major LO haplogroups, with the exclusion of LOb (Extended Data Fig. 1). Using a subset of 461 mitogenomes, including all of the rare lineages, we establish the coalescence times within the LO lineage (Fig. 2a and Supplementary Table 3) and use the complete dataset to reconstruct geographical dispersals (Fig. 2b). We redefine the emergence of the LO lineage to 50–25 thousand years (kyr) before previous estimates<sup>1,6</sup>, around 200 ka (95% confidence interval, 240–165 ka). LOd'k (n = 309; coalesced around 187 ka (the number of mitogenomes and the coalescence time are provided for each lineage)) is largely KhoeSan-specific, emerging approximately 20 kyr before the widely dispersed L0a'b'f'g sister branch (n = 152; around 164 ka). Although the exact branch resolution for LOk remains undetermined, we observe a preference for LOd'k (posterior probability of approximately 0.6) over L0a'b'f'g'k (posterior probability of about 0.4). Irrespective of this, the LOk (n = 113) lineage appears to remain stable for around 130 kyr before diverging into the Kalahari-specific LOk1 lineage, which is predominated by LOk1a (85 out of 94), and rarer LOk1b and LOk2 lineages distributed around the Zambezi river (Extended Data Fig. 2a). The LOd lineage remains stable for almost 60 kyr before splitting into the KhoeSan-specific L0d1'2 and rarer L0d3 lineages.

Coalescing around 113 ka, LOd2 (n = 226) emerges approximately 15 kyr before LOd1 (n = 452). Within LOd2 (emerging about 91 ka), LOd2c diverged the earliest (n = 53; around 84 ka) with a broad and almost even KhoeSan-regional distribution (Extended Data Fig. 3 and Supplementary Table 4). In 2014, we derived an ancient LOd2c1c mitogenome from a sample of the skeleton of an approximately 2,330-year-old Cape-coastal marine forager (St Helena (StHe)/UCT606)<sup>15</sup>. Predating archaeological evidence for sheep herding in the region<sup>12,16</sup>, we proposed that this LOd2c sub-clade represented a pre-pastoral indigenous southern African lineage. Recently, whole-genome sequencing confirmed a unique southern African heritage, whereas two younger (less than 2 kyr old) Cape skeletons showed a genetic link to eastern Africa and the associated pastoralist migration<sup>17</sup>. Previously, an overrepresentation of the LOd2b (28 out of 44; around 65 ka) and LOd2a (62 out of 118; around

60 ka) lineages within the Kalahari KhoeSan has been observed; however, by doubling the contribution of the LOd2d (6 out of 11) lineage, we show a broad southern African distribution (Extended Data Fig. 3 and Supplementary Tables 5, 6). While LOd1 is also spread throughout the KhoeSan-regional identifier, we show notable overrepresentation of the LOd1b (104 out of 174; about 69 ka) and LOd1c (151 out of 184; approximately 59 ka) lineages within the Kalahari and of the LOd1a (32 out of 91; around 44 ka) lineage within the Cape (Extended Data Fig. 4). We contribute two new KhoeSan-ancestral LOd1d mitogenomes to the single published mitogenome.

In contrast to L01'2, the L0d3 lineage is not specific to southern Africa. Although L0d3b (around 30 ka) appears to be KhoeSan-specific, the rarer L0d3a (about 42 ka) lineage is exclusively found north of the Zambezi river. Notably, three out of six L0d3a mitogenomes were derived from east African Sandawe individuals. Our data support previous studies that have suggested a genetic link between east Africa and the earliest southern Africans<sup>17</sup>, who last shared a common ancestor around 59 ka. By adding a large number of mitogenomes (27 out of 40) to the L0d3 lineage, we observe overrepresentation of L0d3b in the Cape KhoeSan identifier (21 out of 34) (Extended Data Fig. 2b and Supplementary Table 7). Using a previously reported identifier that distinguishes maternal KhoeSan ancestry for the Coloured and Baster populations<sup>13</sup>, we show that the L0d3b lineage is specific to the Coloured population, whereas the new L0d2b1a2a sub-clade is specific to the Baster population (Extended Data Fig. 3b).

Within the L0a'b'f'g lineage, L0f is highly divergent (emerging around 125 ka; 95% confidence interval,149–101 ka). By including a further five L0f mitogenomes, we were able to show that L0f1 (13 out of 27; around 113 ka) predominates south and L0f2'3 (14 out of 27; about 121 ka) north of the Zambezi river (Extended Data Fig. 2c and Supplementary Table 8). Within L0f1, we recognize three new branches: the northeast sister clades L0f1c (Zambian) and L0f1b (Tanzanian), and the South African clade L0f1a (n = 8). Lack of L0f representation within contemporary KhoeSan suggests that the presence of L0f1a within South Africa is probably a result of more recent east-coastal agropastoral back-migration. While the L0a'g lineages coalesce around 117 ka (95% confidence interval, 145–94 ka), contributing 19 southern African to 347 L0a mitogenomes, we concur that the L0a lineage probably diverged northeast of the Zambezi river (around 85 ka) and spread throughout Africa<sup>3</sup>; the

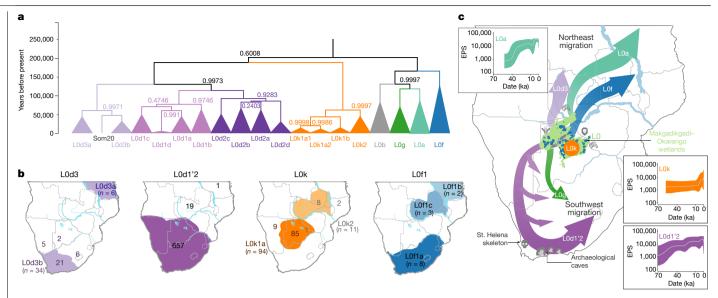


Fig. 2 | LO phylogenetic tree, geographical distributions of the major southern African LO haplogroup and out-of-homeland LO dispersal routes. a, Phylogenetic branching and coalescence times derived from a focused subset of 461L0 mitogenomes, including all rare branches, and anchored to Neanderthals (Homo neanderthalensis: n = 7). The Somalian-derived (Som20) L0d3 mitogenome<sup>3</sup> could not be assigned. **b**, Geographical distribution (identifiers described in Fig. 1b) for all KhoeSan-specific mitogenomes (out of 1,217): L0d3 (n = 40), L0d1'2 (n = 677, excluding one unknown), L0k (n = 105, excluding seven L0k1b and a single Yemen-derived L0k2), and L0f1 (n = 13). Predominant geographical representation (shaded regions), with regionspecific overflow represented by the total number of mitogenomes, including the country-specific representation north of the Zambezi river. c, Schematic map of southern Africa representing the Makgadikgadi-Okavango palaeo-

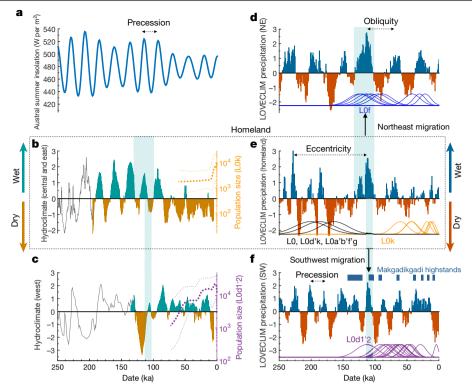
wetland sustained AMH homeland (200-130 ka), supported by archaeological data (represented by the trowel symbol)7 and genetic wildlife data (represented by the lion, zebra and giraffe symbols)<sup>23–25</sup>. The out-of-homeland migration (130-110 ka), results in the split of LOd with LOa'g and LOf divergence. LOd3, LOa and LOf migrate in a northeast direction. LOd1'2 and LOg migrate southwest. while LOk remains in the homeland. Insets show BSP analyses of effective population sizes (EPS) of major LO haplogroups over time, predicting the maintenance of the homeland LOk population (orange), population growth for the broadly dispersed southwest LOd1'2 migrants (purple), which is supported by archaeological evidence (100-60 ka)<sup>8-10</sup> and the StHe mitogenome<sup>15</sup>, while population growth of the northeast LOa migrants coincides with the out-of-Africa migration (aqua). Maps were generated in the R package maps v.3.3.0<sup>37</sup>.

southern representation of the LOa1b and LOa2a lineages are probably a result of a Bantu back-migration (Extended Data Fig. 5). First described in a Kx'a-speaking hunter-gatherer<sup>6</sup>, we now contribute three additional and reclassify five published mitogenomes as LOg (Extended Data Fig. 2d and Supplementary Table 9). As the LOg lineage has a broad KhoeSan and KhoeSan-ancestral distribution, we hypothesize that this lineage diverged southwest of the Zambezi river (around 69 ka), similar to the L0d1'2 lineage.

Our results suggest that the greater Zambezi river basin, particularly the Kalahari region, had a critical role in shaping the emergence and prehistory of AMHs. Now a semi-desert, this region consists of salt pans within northern Botswana that represent desiccated vestiges of palaeo-lake Makgadikgadi, which at its peak in the early Pleistocene would have been the largest lake in Africa<sup>7,18</sup>. Contraction of the Makgadikgadi palaeo-lake during the Middle Pleistocene was accompanied by development of the Okavango delta as a result of neotectonic rifting, which—together with smaller lakes from the upper Zambezi to the Kafue rivers-would have created a vast residual wetland favourable for habitation by humans and mammals more broadly<sup>19</sup> (Fig. 2c). Today, the harsh Kalahari climate and oxygen-rich salt pans are not ideal for fossil and pollen preservation, respectively. However, period-relevant lithic artefacts are documented from the Makgadikgadi pans and surroundings<sup>7,20,21</sup>, while palynology suggests that this region was once a grassland and forest biome<sup>22</sup>. Our data further suggest that the Makgadikgadi-Okavango palaeo-wetland sustained the existence of AMHs for around 70 kyr, supported by mitochondrial data of ancestral giraffe, lion and zebra<sup>23-25</sup>, before out-of-homeland migrations split the founder homeland populations of the LOd, LOf and LOa'g lineages.

Southwest of their homeland, the L0d1'2 lineage experienced episodic splits and showed a broad south-coastal occupation of the emerged sub-populations, whereas the ancestors of the LOg lineage were less successful. Bayesian skyline plot (BSP) (Fig. 2c) analysis confirms effective population growth for the LOd1'2 lineage (BSP LOd1'2), whereas extensive archaeological evidence indicates cognitively modern human behaviour at the southern tip of Africa<sup>8-10</sup> between approximately 100 and 60 ka, together with an associated increase in the density of timeappropriate archaeological sites in coastal compared to inland regions<sup>26</sup>. Northeast of their homeland, the LOd3 and LOf lineages are less successful, whereas the LOa lineage underwent considerable diversification. which post-dates the out-of-Africa migration (BSP LOa; Fig. 2c). The northeast migration route is further supported by the appearance of data-appropriate archaeological sites<sup>26</sup>. Within their homeland, the  $population\, carrying\, the\, LOk\, lineage\, sustained\, a\, constant\, effective\, population\, carrying\, the\, LOk\, lineage\, sustained\, a\, constant\, effective\, population\, carrying\, the\, LOk\, lineage\, sustained\, a\, constant\, effective\, population\, carrying\, the\, LOk\, lineage\, sustained\, a\, constant\, effective\, population\, carrying\, the\, LOk\, lineage\, sustained\, a\, constant\, effective\, population\, carrying\, the\, LOk\, lineage\, sustained\, a\, constant\, effective\, population\, carrying\, the\, LOk\, lineage\, sustained\, a\, constant\, effective\, population\, carrying\, the\, LOk\, lineage\, sustained\, a\, constant\, effective\, population\, carrying\, the\, LOk\, lineage\, sustained\, a\, constant\, effective\, population\, carrying\, the\, LOk\, lineage\, sustained\, a\, constant\, effective\, population\, carrying\, ca$ lation size (BSPLOk), as did the Kalahari-predominant LOd2b, LOd2a and LOd1c lineages. Although the presence of LOk in Zambia has been suggested to represent contact with an ancient pre-Bantu population<sup>27</sup>, we propose that these rare lineages represent an ancient out-of-homeland branch of the ancestral KhoeSan population.

Orbitally driven large-scale hydroclimate variations have been proposed as a contributor of early human migrations <sup>28,29</sup>. In some studies, wetter conditions and resulting 'green corridors' have been proposed to explain the out-of-Africa migration (a 'pull' scenario), whereas others have proposed that drier conditions and resulting food shortages forced dispersals (a 'push' scenario)<sup>30</sup>. To determine whether our predicted homeland isolation and major dispersals may have been driven by climate shifts, we analysed four key palaeo-hydroclimate datasets 29,31-33, along with a transient 784-kyr-long glacial-interglacial simulation conducted with the LOVECLIM Earth system model<sup>28</sup> (Fig. 3). Although limited by available palaeo-proxy records and a climate model of intermediate complexity, we observe a considerable degree of coherence on orbital timescales (Extended Data Fig. 6). During the homeland period (200-130 ka), palaeo-data link the 21-kyr-long precession cycle, which



**Fig. 3** | **Reconstructed and simulated climatic conditions during the out-of-homeland migration. a**, Austral summer insolation changes (blue) at  $27^{\circ}$  S. **b**, A hydroclimate composite of eastern and central southern Africa (shading) was obtained by averaging the Fe/K runoff record from core CD154-1006P³¹ and the Pretoria Salt Pan rainfall reconstruction²°, extended from 250 to 190 ka (grey line). The plot shows the effective population size for homeland L0k as analysed by BSP (orange dashed lines). **c**, Southwestern hydroclimate reconstruction (shading) obtained by averaging normalized leaf wax data (MD08-3167)³³ and the aridity index from cores (MD96-2094)³², for which the aridity record

extended from 250 to 140 ka (grey line) and the effective population size of L0d1'2 as analysed by BSP is shown (purple dashed lines).  $\bf d$ , Simulated LOVECLIM normalized precipitation changes (shading) northeast of the homeland (33° E, 13° S) and coalescence time probabilities for L0f haplogroup (blue bell curves).  $\bf e$ , Same as for  $\bf d$ , but for the homeland and coalescence probabilities for L0, L0d'k, L0a'b'f'g (black) and L0k haplogroups (orange).  $\bf f$ , Same as for  $\bf d$ , but for the area southwest of the homeland (17° E, 30° S) and L0d1'2 coalescence times (purple). Blue bars indicate predicted Makgadikgadi high stand phases 35. NE, northeast; SW, southwest.

arises from a combination of Earth's axis wobble and a slow rotation of Earth's entire orbit around the Sun (Fig. 3a), with three wet–dry cycles (Fig. 3b). By contrast, the climate model simulates an extended drought, owing to a more pronounced eccentricity signal (Fig. 3e), suggestive of a wetland oasis in an otherwise vast harsh environment.

During the out-of-homeland period (130–110 ka), our model simulation supports humid conditions to the northeast that facilitated the first dispersals, concurring with LOf coalescence (around 125 ka) (Fig. 3d). By contrast, the region southwest of the homeland experienced an approximately 15-kyr-long megadrought before an orbital shift created the favourable humid conditions that led to the dispersal of the LOd1'2 lineage (around 113 ka) (Fig. 3f), which is also supported by palaeo-data (Fig. 3c). This is also around the time the northeast LOa and southwest LOg migrants last share a common ancestor (around 117 ka). During the last glacial period (approximately 100-11 ka), we observe a reduction in the amplitude of the changes in orbital-scale hydroclimate and overall drying within the homeland (Fig. 3b), whereas the southwest coastal hydroclimate was dominated by precessional variability and showed relatively agreeable environmental conditions (Fig. 3c, f). Notably, periods of deceleration and acceleration in the estimates of the effective population size of the LOd1'2 lineage coincide with regional changes in hydroclimate, further linking climate, population size and evolution.

We propose that the Makgadikgadi–Okavango palaeo-wetland was the possible homeland of AMHs. Although one cannot exclude the possibility of a polycentric origin  $^{34}$ , this deltaic–lacustrine ecosystem would have provided an ideal geographical locality for the evolution and 70-kyr-long sustained existence of the deepest-branching maternal founder population of AMHs. Increased humid conditions, supported by

palaeo-lake system reconstructions<sup>35</sup>, between 130 and 110 ka would have opened green corridors for successful northeast-southwest migrations, supporting a pull scenario. Drying within the homeland following the out-of-homeland period, supported by hydroclimate data (110–100 ka) and a model simulation (100–80 ka), would have created a push scenario. in which a reduced carrying capacity of the land would have increased pressure to seek out climatically more favourable regions. We propose that the southwest migrants maintained a successful coastal forager existence, while the northeast migrants—similar to the later-branching population of L1'6-gave rise to ancestral pastoral and farming populations. A recent publication<sup>36</sup> provides further mitochondrial evidence to support the northeast out-of-homeland migration route and expansion into eastern Africa around 70-60 ka. Revealing a southern African homeland for the emergence and extended subsistence of the LO lineage,we propose that an out-of-homeland migration event, which was probably driven by astronomically induced regional shifts in hydroclimate, shaped the present-day ethnic and genetic diversity of modern humans.

### **Online content**

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1714-1.

 Behar, D. M. et al. The dawn of human matrilineal diversity. Am. J. Hum. Genet. 82, 1130–1140 (2008).

- Brown, F. H., McDougall, I. & Fleagle, J. G. Correlation of the KHS tuff of the Kibish Formation to volcanic ash layers at other sites, and the age of early Homo sapiens (Omo I and Omo II). J. Hum. Evol. 63, 577-585 (2012).
- Rito, T. et al. The first modern human dispersals across Africa. PLoS ONE 8, e80031 (2013).
- Stringer, C. & Galway-Witham, J. On the origin of our species. Nature 546, 212-214 (2017).
- Henn, B. M. et al. Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. Proc. Natl Acad. Sci. USA 108, 5154-5162 (2011).
- Chan, E. K. F. et al. Revised timeline and distribution of the earliest diverged human maternal lineages in southern Africa. PLoS ONE 10, e0121223 (2015).
- Moore, A. E., Cotterill, F. P. D. & Eckardt, F. D. The evolution and ages of Makgadikgadi palaeo-lakes: consilient evidence from Kalahari drainage evolution south-central Africa. S. Afr. J. Geol. 115, 385-413 (2012).
- Henshilwood, C. S. et al. A 100,000-year-old ochre-processing workshop at Blombos 8. Cave. South Africa. Science 334, 219-222 (2011).
- Douze, K., Wurz, S. & Henshilwood, C. S. Techno-cultural characterization of the MIS 5 9 (c. 105-90 ka) lithic industries at Blombos cave. Southern Cape. South Africa, PLoS ONE 10. e0142151 (2015).
- Henshilwood, C. S. et al. An abstract drawing from the 73.000-year-old levels at Blombos Cave. South Africa. Nature 562, 115-118 (2018).
- Güldemann, T. in Beyond 'Khoisan': historical relations in the Kalahari Basin (Current Issues in Linguistic Theory 330) (eds Güldemann, T. & Fehn, A.-M.) 330, 1-40 (John Benjamins, 2014)
- 12 Lander, F. & Russell, T. The archaeological evidence for the appearance of pastoralism and farming in southern Africa. PLoS ONE 13, e0198941 (2018).
- Petersen, D. C. et al. Complex patterns of genomic admixture within southern Africa. PLoS Genet. 9, e1003309 (2013).
- Morris, A. G. Isolation and the origin of the Khoisan: late Pleistocene and early Holocene human evolution at the southern end of Africa. Hum. Evol. 17, 231-240 (2002).
- Morris, A. G., Heinze, A., Chan, E. K. F., Smith, A. B. & Hayes, V. M. First ancient mitochondrial human genome from a prepastoralist southern African. Genome Biol. Evol. 6, 2647-2653 (2014)
- Pleurdeau, D. et al. "Of sheep and men": earliest direct evidence of caprine domestication in southern Africa at Leopard Cave (Erongo, Namibia). PLoS ONE 7, e40340 (2012).
- 17. Skoglund, P. et al. Reconstructing prehistoric African population structure. Cell 171, 59-71
- Eckardt, F. D. et al. Mapping the surface geomorphology of the Makgadikgadi Rift Zone (MRZ). Quat. Int. 404, 115-120 (2016).
- Wrangham, R. W. in Interpreting the Past: Essays on Humans, Primates and Mammal 19 Evolution (eds Pilbeam, D. R. et al.) 231-242 (Brill Academic, 2005).
- Robbins, L. H. et al. The advent of herding in southern Africa; early AMS dates on domestic livestock from the Kalahari Desert, Curr. Anthropol. 46, 671-677 (2005).
- Mackay, A., Stewart, B. A. & Chase, B. M. Coalescence and fragmentation in the late 21. Pleistocene archaeology of southernmost Africa. J. Hum. Evol. 72, 26-51 (2014).

- Scott, L. & Neumann, F. H. Pollen-interpreted palaeoenvironments associated with the Middle and Late Pleistocene peopling of Southern Africa. Quat. Int. 495, 169-184 (2018).
- 23. Bock, F. et al. Mitochondrial sequences reveal a clear separation between Angolan and South African giraffe along a cryptic rift valley. BMC Evol. Biol. 14, 219 (2014).
- Pedersen, C. T. et al. A southern African origin and cryptic structure in the highly mobile Plains zebra. Nat. Ecol. Evol. 2, 491-498 (2018).
- Moore, A. E. et al. Genetic evidence for contrasting wetland and savannah habitat specializations in different populations of lions (Panthera leo), J. Hered. 107, 101-103
- Blome, M. W., Cohen, A. S., Tryon, C. A., Brooks, A. S. & Russell, J. The environmental context for the origins of modern human diversity: a synthesis of regional variability in African climate 150,000-30,000 years ago. J. Hum. Evol. 62, 563-592 (2012).
- Barbieri, C. et al. Ancient substructure in early mtDNA lineages of southern Africa. Am. J. Hum. Genet. 92, 285-292 (2013).
- Timmermann, A. & Friedrich, T. Late Pleistocene climate drivers of early human migration. 28 Nature 538, 92-95 (2016).
- Partridge, T. C., Demenocal, P. B., Lorentz, S. A., Paiker, M. J. & Vogel, J. C. Orbital forcing 29 of climate over South Africa: a 200,000-year rainfall record from the Pretoria saltpan. Quat. Sci. Rev. 16, 1125-1133 (1997).
- 30. Tierney, J. E., deMenocal, P. B. & Zander, P. D. A climatic context for the out-of-Africa migration, Geology 45, 1023-1026 (2017).
- Simon, M. H. et al. Eastern South African hydroclimate over the past 270,000 years. Sci. Rep. 5, 18153 (2015).
- Stuut, J.-B. W. et al. A 300-kyr record of aridity and wind strength in southwestern Africa: inferences from grain-size distributions of sediments on Walvis Ridge, SE Atlantic. Mar. Geol. 180, 221-233 (2002).
- Collins, J. A., Schefuß, E., Govin, A., Mulitza, S. & Tiedemann, R. Insolation and glacialinterglacial control on southwestern African hydroclimate over the past 140 000 years. Earth Planet. Sci. Lett. 398, 1-10 (2014).
- Scerri, E. M. L. et al. Did our species evolve in subdivided populations across Africa, and why does it matter? Trends Ecol. Evol. 33, 582-594 (2018).
- Burrough, S. L., Thomas, D. S. G. & Bailey, R. M. Mega-lake in the Kalahari: a late Pleistoscene record of the palaeolake Makgadikgadi system. Quat. Sci. Rev. 28, 1392-1411
- Rito, T. et al. A dispersal of Homo sapiens from southern to eastern Africa immediately preceded the out-of-Africa migration. Sci. Rep. 9, 4728 (2019).
- Becker, R. A. & Wilks, A. R. maps: Draw Geographical Maps. R package version 3.3.0 https://cran.r-project.org/web/packages/maps/index.html (2018).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

# Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

### Statement on population identifiers

The authors acknowledge that population identifiers (or ethnic labels) have different meanings to different peoples across different countries and between and within different ethnic groups. During the apartheid rule, South Africans were grouped according to ethnic identities, which resulted in discrimination based on population identifiers such as Bantu or Coloured. In turn, others view the very same population identifiers with cultural identity and pride. In 2013, we performed a study led by a Coloured co-author to assess the sensitivity in self-identification as Coloured, Of 521 participants, 91.2% self-identified as Coloured, Cape Coloured or South African Coloured, while 8.8% elected against the use of Coloured for self-identification<sup>14</sup>. In turn, using such population identifiers within the context of the United States would be seen as derogatory and highly offensive. We have previously genetically profiled the Baster population of Namibia<sup>13</sup> and again what could be to others a derogatory term, to the Baster community of Rehoboth in Namibia, the term is used with immense pride, who recognize themselves as a Republic with a national flag<sup>38</sup>.

In this study, the authors have used linguistics, supported by ethnicity, to provide population identification, with further historical, geographical and genetic classification for deriving maternal contributions (described in the next section). KhoeSan (or KhoeSaan) languages are grouped together due to their use of click consonants as a unique language identifier. Once spread across the entire southern African region, KhoeSan languages are today restricted largely to populations residing in Namibia and Botswana (and southern Angola), although two Tanzanian isolates, Sandawe and Hadza, are believed to be linguistically related click languages (or east African KhoeSan)<sup>39</sup>. 'San' literally means 'forager' and Khoe means 'person'; culturally, the KhoiSan identifier refers to hunter-foragers (San) or herders (Khoi). At times linguistic and cultural identities clash. For example, Nama and Hai||om peoples both speak Nama (a Khoe-Kwadi language), while culturally and historically these two populations are quite different, representing a herder and hunter-gatherer ancestry, respectively. Additionally, autosomal genetic data have been used to provide further insights into KhoeSan admixture and substructures, highlighting at a genetic level the historical differences between the Nama and Hai||om<sup>40</sup>. We have attempted to capture both ethnic and linguistic identifiers that best reflect population ancestry. In contrast to KhoeSan languages, most Bantu languages do not contain click consonants; however, exceptions exist within Southern African Bantu languages (for example, isiXhosa and isiZulu languages, which have borrowed click consonants from their KhoeSan neighbours). Spoken across the entire sub-Saharan Africa (up to 500 groups), the Guthrie classification of languages further identifies the S-zone or Southern Bantu (South Africa, Zimbabwe, southern Mozambique and most of Botswana) and the R-zone or Southwest Bantu languages (northern Namibia, southern Angola and northwest Botswana)41, which are of relevance to this study.

### Ethics statement and recruitment

The study was performed in accordance with the ethical standards of the overseeing human research ethics committees and local governance, as per the 1964 Helsinki Declaration. The study was reviewed and approved by the Ministry of Health and Social Services (MoHSS) in Namibia (17-3-3 2008, 2014 and 2019), with additional local approvals from participating community leaders, the University of Pretoria Human Research Ethics Committee (HREC 43/2010 and HREC 280/2017), including US Federal-wide assurance (FWA00002567 and IRB00002235 IORG0001762), as well as the South African National

Blood Service (SANBS) HREC (HREC 2012/11). Participants were recruited within the borders of Namibia and South Africa and self-reported ethno-linguistic population identifiers were recorded. Blood samples were taken after receiving written and/or recorded informed consent. Isolated DNA was shipped under the Republic of South Africa Department of Health Export Permit (J1/2/4/2), in accordance with the National Health Act 2003, to the Garvan Institute of Medical Research in Australia. Mitogenome sequencing was performed in accordance with site-specific approval granted by St Vincent's Hospital HREC in Australia (SVH 15/227).

### Participant population identifiers

Merging with published data for a total of 1,217 L0 mitogenomes, participants were broadly classified as KhoeSan, Bantu or Cape multiethnic heritage. Indigenous KhoeSan who inhabit the inland semi-desert Kalahari region of Botswana and Namibia include the Kx'a (Jul'hoan or Hoan, and !Xun or !Xuun), Tuu (or Taa) and Khoe-Kwadi (Naro, ||Ani, Khwe, Buga, G||ana, G||ui, ||Xokhoe, Tshwa and Shua) speakers. Indigenous KhoeSan who inhabit the west-coastal region of Namibia speak a Khoe-Kwadi or Nama language and include the Nama, Damara, Topnaar (‡Aonin) and Hai||om speakers<sup>42,43</sup>. Novel mitogenomes were derived from 15 Kalahari KhoeSan, including Ju|'hoan (n=9), !Xun (n=1)and Naro (n=5), and 21 west-coastal KhoeSan, including Nama (n=7), Damara (n = 8) and Topnaar ( $\ddagger$ Aonin, n = 6) from Namibia. Speakers of southwest Bantu (non-KhoeSan) languages (which do not contain click consonants) of Namibia, Botswana and southerly boarders of Angola, presenting with KhoeSan-predominant LO maternal lineages, most likely carry a Kalahari or west-coastal KhoeSan mitogenome. As a result of refuge provided to the Herero by the Kalahari KhoeSan during the early 1900 German South West African genocide<sup>44</sup>, we speculate in this study a probable Kalahari KhoeSan heritage for the three Herero mitogenomes.

 $Although in digenous \, Khoe San\, are \, arguably \, absent \, from \, the \, coastal$ regions of South Africa, and while recognising and honouring the northwest inland (southern Kalahari) †Khomani San of South Africa (although not recruited within the context of this study), KhoeSan skeletal remains spread across the region<sup>45</sup>. Hunter-gatherer KhoeSan once inhabited a broad southwest to east-coastal region at the tip of Africa. These skeletal remains predate archaeological evidence supporting the arrival of sheep herders who appear to have crossed the Okavango river in northern Namibia around 2.2 ka, migrating along the southwest coast to the southern Cape<sup>12,16,20,45</sup> by around 2 ka. Recently, Cape KhoeSan skeletons younger than 2 ka have been genetically linked to east Africa and herder migration<sup>17</sup>. Migrating herders may have acquired indigenous KhoeSan maternal contributions. Along the east coast, southwardmigrating Bantu farmers (Southern Bantu; who presumably did not speak languages containing click consonants) entered South Africa around 1,500 years ago, while a second wave of Bantu migrants (Southwest Bantu) crossed central Africa into Namibia around 800 years ago<sup>12</sup>. Maternal contributions to the South African Southern-Bantu-speaking populations (n = 43, this study) may therefore either be of Bantu origin (in this case, LOa lineages and therefore non-KhoeSan) or of east-coastal KhoeSan-ancestry. The arrival of European colonists and Dutch-East-Indian slaves to the Cape in the mid-1600s, gave rise to a multi-ethnic (European, Asian, KhoeSan and Bantu) Cape population, the ancestors of the South African Coloured (n = 90, this study) and Namibian Basters (n = 24, this study), who historically speak a Dutch-derived language  $known\,as\,Afrikaans^{13,46}. Emerging\,from\,a\,common\,historical\,background$ to the Coloured, the Baster population have since the late 1800s distinguished themselves as independent from the Coloured, migrating to the Baster nation of Rehoboth in Namibia<sup>47</sup>. Although the vast majority of LO mitogenomes represented in the Baster and Coloured populations are of Cape KhoeSan heritage (100% and 94.4%, respectively), we observe a percentage of non-KhoeSan (Bantu) LOa lineages within the Coloured population.

### L0 haplogroup pre-screening

Subjects were selected for whole-mitogenome sequencing based on pre-screening for specific LO markers using direct amplicon-specific Sanger sequencing. Specifically, a 2,673-bp region (Cambridge Reference Sequence (rCRS) position 3322–5995) was amplified and initially screened for the LO variant T5442C. LO samples were further screened to delineate LOd (T4232C), LOd1 (G3438A), LOd1b (T3618C), LOd1c (C4197T), LOd1'2 (A3756G), LOd2 (A3981G, C205T, A4044G), LOd2a (A5153G), LOd2d (G5147A, G5231A), LOd2c (A4038G, T4937C) and LOd3 (G5460A, G5773A) lineages. This identified 188 samples carrying a rare LO haplogroup: LOd1b (n=21), LOd1c (n=13), LOd2a (n=30), LOd2b (n=7), LOd2c (n=15), LOd2d (n=6), LOd3 (n=29), LOa1 (n=6), LOa2 (n=6), LOf (n=5) and LOk (n=5); as well as 55 samples that could not be unambiguously assigned to a major LO sub-lineage: LOd1a'c (n=2), LOa'b'f'k (n=5), LOa'b (n=2), LOd2 (n=1) and LOd1 (n=45, assumed LOd1a) (Supplementary Table 1).

### Whole-mitogenome sequencing

Mitogenomes were isolated using two overlapping amplicons as previously described<sup>6,48</sup>. Specifically, two primer pairs were used to isolate and amplify fragments 12,250-3,005 (7.2 kb) and 2,583-12,337 (9.7 kb) of the circular mitogenome. This pair of primers has been demonstrated to effectively capture the mitogenome with high specificity while minimizing off-target capture of nuclear copies of mitochondrial-derived DNA. Following touchdown long-range amplification with the Platinum Taq DNA Polymerase High Fidelity (Invitrogen), the two amplicons were purified using AMPure XP beads (Agencourt) and combined in a 7:13 ratio of short:long fragments. Sequencing was performed on the Ion Torrent PGM platform. In brief, 200-bp single-end sequencing libraries were prepared using the Ion Xpress Plus Fragment Kit and Ion Xpress Barcode Adaptors (ThermoFisher), and 4-16 samples (barcodes) were pooled and sequenced on 314v2 Ion Chips. Using the Ion Torrent suite v.5.0.2.1, sequencing reads were quality trimmed and aligned to the human mitochondrial revised rCRS (accession NC\_012920.1). Consensus mitogenome sequences were derived by first identifying variants relative to rCRS, using samtools (v.1.3.1) mpileup (with parameters -d 10000 -L 1000 -Q7-h50-o10-e17-m4)<sup>49</sup> and beftools (v.1.3.1) call (with parameters-c-M) (http://www.htslib.org/doc/bcftools.html), then converting to the FASTA format using the vcfutils.pl vcf2fq program in samtools.

### Publicly available data

An exhaustive search for publicly available L0 mitogenomes was performed between 2015 and 2017, identifying 26 studies comprising a total of 6,334 mitogenomes. L0 status for all mitogenomes was deduced, either directly from the original publication or by downloading the nucleotide sequences from NCBI and evaluating their haplogroup using HaploGrep2 (v.2.1.13) $^{50}$  based on PhyloTree Build  $17^{51}$ . From this dataset, a subset of 1,019 L0 mitogenomes was identified and included in this study (Extended Data Table 1 and Supplementary Table 2). Publicly available genomes were broadly classified as KhoeSan, Bantu (KhoeSan ancestral), or non-KhoeSan based on the reported population and/or country of origin.

# Whole-mitogenome haplotyping

HaploGrep2<sup>50</sup> was used to type all 1,217 sequences against PhyloTree Build  $17^{51}$ . This resulted in the refinement and reclassification of our 198 mitogenomes, resulting in LOd1 (n = 81, including 45 LOd1a, 21 LOd1b, 13 LOd1c and 2 LOd1d), LOd2 (n = 58, including 30 LOd2a, 8 LOd2b, 14 LOd2c and 6 LOd2d), LOd3 (n = 27), LOa (n = 19), LOf (n = 5), LOk (n = 5) and LOg (n = 3) mitogenomes (Supplementary Table 1). This refined, and in some cases reclassified, the haplogroups of the 1,019 publicly available mitogenomes (Supplementary Table 2).

# Phylogenetic inference

Multiple sequence alignment was performed across all 1,217 mitogenomes along with 7 Neanderthal genomes (Supplementary Table 10),

using MUSCLE v.3.8.31 $^{52}$  with parameters -maxiters 3 -diags1. Phylogenetic inference was performed using FastTree v.2.1.7 (SSE3) $^{53}$  using the generalized time reversible (-gtr) and discrete gamma model with 20 rate categories (-gamma). A summary of the inferred phylogenetic tree is shown in Extended Data Fig. 1, with the tree rerooted to the 7 Neanderthal genomes.

Bayesian phylogenetic inferences and divergence times were calculated using BEAST2 v.2.4.2 with BEAGLE v.2.0<sup>54</sup>. Owing to the computational burden of this analysis, BEAST was performed on a subset of 461 mitogenomes, selected to include: (i) only complete mitogenomes (27 mitogenomes with only the coding region<sup>55,56</sup> were excluded); (ii) all 198 novel mitogenomes from this study; (iii) all 121 L0 mitogenomes from our previous studies, Chan et al.<sup>6</sup> (n = 77), Morris et al.<sup>15</sup> (StHe, defining the new haplogroup L0d2c1c), McCrow et al.<sup>48</sup> (n = 37) and Schuster et al.<sup>57</sup> (n = 6); (iv) all rare haplogroups, namely L0g (n = 9), L0f (n = 22), L0d3 (n = 30), L0d1d (n = 3), L0d2d (n = 11) and L0k2 (n = 12); (v) all mitogenomes that could not be unambiguously typed by HaploGrep2<sup>50</sup> (n = 14; none from this study); and (vi) a random subset of mitogenomes for all remaining sub-lineages not already represented.

Multiple sequence alignment of the subset of 461 AMH and 7 Neanderthal mitogenomes was converted to NEXUS format using the convert function of seqmagick v.0.6.1 (https://fhcrc.github.io/seqmagick) with parameter -- alphabet dna-ambiguous. This provided the input to BEAST2. Specifically, BEAUTi v.2.4.2 was used to set up the phylogenetic model, assuming: (i) the gamma site model with six gamma categories and no invariant sites; (ii) the generalized time-reversible substitution model; (iii) a strict constant clock model with a normal prior with  $\mu = 1.665 \times 10^{-8}$  and  $\sigma = 1.479 \times 10^{-9}$  based on a previously published study<sup>58</sup>; and (iv) a coalescent constant population. Times were calibrated to the seven H. neanderthalensis mitogenomes with tip dates set to their reported approximate archaeological dating estimates: Feldhofer 1, 40 ka<sup>59</sup>; Vindija, 38 ka<sup>59</sup>; El Sidron, 39 ka<sup>59</sup>; Feldhofer 2, 40 ka<sup>59</sup>; Mezmaiskaya, 65 ka<sup>59</sup>; Croatia, 38.31 ka<sup>60</sup>; Altai, 50 ka<sup>61</sup> (Supplementary Table 10). No prior was set on the most recent common ancestor of this taxon set, and calibration was applied to the leaves instead of the most recent common ancestor. Further, a normal prior,  $N(\mu = 200,000, \sigma = 50,000)$ , was set on the coalescent time of the AMH genomes, and a tip date of 2,330 years before present was set for the StHe genome<sup>15</sup>.

Five BEAST replicates were performed, each with 100 million Markov chain Monte Carlo iterations, sampling every 10,000. Tracer v.1.6 was used to evaluate BEAST trace files (Supplementary Table 11), ensuring all runs had converged. The five replicates were combined using LogCombiner v.2.4.2, discarding 10% of the samples as burn-in for each replicate and without resampling states at a lower frequency.

Sampled trees from BEAST were summarized into a single maximum clade credibility target tree using TreeAnnotator v.2.4.2 for each of the five replicates, discarding the first 10% as burn-in. To summarize across replicates, sampled trees from the five replicates were first combined using LogCombiner v.2.4.2, again discarding the first 10% as burn-in from each replicate, but resampling at a lower frequency of 50,000 (five replicates of 10,000 samples). The combined, resampled trees were then summarized with TreeAnnotator v.2.4.2 as for the individual replicate BEAST results.

FigTree v.1.4.2+ (http://tree.bio.ed.ac.uk/software/figtree/) was used to visualize all resulting trees.

### **BSP** analysis

BSP analyses were performed to estimate the demographic history of each maternal haplogroup. Although maternal haplogroups do not necessarily equate to population groups, it has been suggested that the signal associated with a haplogroup can still provide insights into the demographic processes in the populations who carry the haplogroup 62,63.

For each haplogroup of interest (for example, L0a, L0d1'2 and L0k), a nexus file was derived using SeqMagic v.0.6.1 as described above.

BSP analyses were performed using BEAST2, using BEAUTi2 for model setup as described in 'Phylogenetic inference', with the following key differences: (i) the gamma shape of the gamma site model was estimated with an exponential prior with mean = 1.0 and offset = 0.0; (ii) the molecular clock was fixed (not estimated) at  $1.665 \times 10^{-8}$  based on a previously published study<sup>58</sup>; and (iii) the phylogenetic tree prior was set to coalescent Bayesian skyline, assuming 20 intervals between the root of the tree and the present time.

Tracer v.1.6 was used to reconstruct the Bayesian skyline from the sampled trees for each analysis, using a stepwise constant variant and the lower 95% highest posterior density of the root height as the maximum time. Results of this analysis are summarized in Supplementary Table 12.

### Geographical history of the palaeo-wetland Makgadigadi

Initiated around 2 million years ago, palaeo-lake Makgadikgadi $^7$  originally covered an area of around 170,000 km² at its highest lake stand, bounded by a shoreline of around 995 m. A degraded sand ridge (the Deception ridge), was associated with the 995-m shore in the southwest of the lake. This lake would have covered more than twice the area of modern Lake Victoria, and similar to the latter, would have caused a considerable climatic feedback, with locally enhanced rainfall. We previously proposed that this was, in turn, responsible for the initiation of the surrounding (now-fossil) drainages, creating a well-watered environment and very favourable habitat for mammals, including hominids $^7$ . Smaller lakes, now represented by residual wetlands, also formed on the upper Zambezi river and the modern Kafue Flats on the Kafue river, resulting in an archipelago of palaeo-lakes in south-central Africa during the Early and Middle Pleistocene epoch.

Palaeo-Makgadikgadi, bounded by the 995-m shoreline, was originally sustained by a major drainage line, which included the Chambeshi river as headwaters, connected to the upper Zambezi river via the upper Kafue river. Severance of the original links between the Chambeshi river and upper Kafue river, and the latter and the upper Zambezi river resulted in a sequential contraction of the Makgadikgadi to a much smaller water body. This is reflected in a series of fossil shorelines, associated with breaks in slope, at progressively lower levels (945 m, 936 m and 922 m). The Gidikwe ridge was associated with the 945-m shoreline. However, contraction of the lake was accompanied by the development of the modern Okavango delta. Timing of the contraction of the lake and initiation of the Okavango delta is not tightly constrained, but by the time that we propose that modern humans emerge within the region. around 200 ka, we speculate that the formerly extensive Makgadikgadi palaeo-lake had contracted to a much less extensive deltaic-lacustrine system. Together with the lakes that developed from the upper Zambezi and Kafue rivers to the north and the Okavango delta to the west, the region would have been a vast wetland, a favourable habitat for hominid occupation. It is this palaeo-wetland region that we propose as the homeland for the founder population of AMHs.

### Climate model simulations and palaeo-climate data

To place the coalescence time estimates of the LO branch into a climatic context and to test the robustness of simulated hydroclimate responses in South Africa to orbital-scale conditions, we use the LOVE-CLIM Earth system model of intermediate complexity  $^{28}$ . It is based on a 3-layer atmosphere, a 20-level ocean general circulation model, a dynamic–thermodynamic sea–ice model and a terrestrial vegetation model. A transient simulation that covers several glacial–interglacial cycles was conducted using time-dependent boundary conditions. The experiment  $^{28}$  (covering the past 784 kyr) uses time-varying boundary conditions for orbital parameters,  $\mathrm{CO}_2$  and other greenhouse gas concentrations obtained from Antarctic ice cores, and an estimate of Northern Hemispheric ice-sheet orography and albedo changes (data are used in Fig. 3 and Extended Data Fig. 6). The forcings are applied with an acceleration factor of five: one coupled model year corresponds to five orbital calendar years. Our analysis focuses on the past 250 kyr in

both simulations. The climate sensitivity of this model to  $CO_2$  variations was modified to capture the range of reconstructed global mean surface temperature changes in response to radiative forcing  $^{64}$ . The transient LOVECLIM model simulations have previously been validated against other palaeo-climate records from around the world  $^{28,64,65}$ . Our analysis here focuses on the simulated precipitation as well as changes in tree and grass fractions in central eastern Africa and western southern Africa (data used in Fig. 3 d–f and Extended Data Fig. 6b–d).

As a result of its coarse horizontal atmospheric resolution  $(5.6^\circ)$  and the use of only parameterized ageostrophic wind components, LOVE-CLIM has several deficiencies. Of particular note are the lack of realistic El Niño–Southern Oscillation variability and the fact that annual mean freshwater flux corrections have been applied to mimic the atmospheric moisture transport from the Atlantic to the Pacific and to stabilize the Atlantic Meridional Overturning Circulation.

There exist only a few long-term hydroclimate datasets from southern Africa that cover the past >120 kyr. Here we compare the simulated LOVE-CLIM precipitation (normalized) in central southern Africa with a southern central African hydroclimate composite, obtained by averaging the normalized orbitally tuned rainfall reconstruction from the Pretoria salt pan<sup>29</sup> and the normalized Fe/K river run-off proxy obtained from marine sediment core CD154-1006P31 (Fig. 3b). The composite index emphasizes the joint variability in both records. We find that some of the overall features in the observations-particularly the fact that rainfall is modulated by the precessional cycle of austral summer insolation<sup>66</sup> (Fig. 3a)—are well-captured by the LOVECLIM model experiment. However, we also find some discrepancies in the central part of southern Africa, such as in the phase of the precessional signal and the difference in overall wet and dry conditions during the homeland period from 200 to 120 ka. The overall glacial drying in the central part of southern Africa from 100 to 20 kais, however, captured in both model simulation and palaeo-proxy reconstructions (Fig. 3b, e). Orbital-scale hydroclimate variations in southern Africa are clearly not spatially homogenous (Fig. 3b-f). To gain a better understanding of the spatial patterns of hydroclimate variability, we compared the model simulation with a composite index from southwestern Africa, obtained by averaging a normalized aridity index reconstructed from sediment core MD96-2094<sup>33</sup> and the normalized  $\delta^{13}$ C isotope ratio data of leaf wax extracted from the South Atlantic sediment core MD08-3167<sup>32</sup> (Fig. 3c and Extended Data Fig. 6c, d). The results show a good correspondence between model and reconstructions on the western side of southern Africa, and in particular reproduce a major drought period that peaked around 120 ka and a subsequent increase in rainfall towards the last glacial period. This gradual increase in rainfall corresponds to an overall increase in lineage splitting of the LOd1'2 haplogroup (Fig. 3f) and growth of its population (Fig. 3c). This result further highlights the possibility that climate shifts may have played an important part in the southwestward migration of LOd1'2 descendants (Fig. 2).

To further test the fidelity of LOVECLIM in reproducing interhemispheric orbital rainfall shifts across Africa, we also compared the simulated vegetation changes with a leaf-wax index from stable hydrogen isotope data extracted from a sediment core in the Gulf of Aden<sup>30</sup>, which is indicative of hydroclimate and vegetation changes in the northeastern Horn of Africa (Extended Data Fig. 6b). The comparison shows a good qualitative correspondence for the precessional-scale timing of rainfall and vegetation maxima and minima as well as of the eccentricity modulated amplitude of these changes, lending further support to the credibility of the simulated rainfall patterns across Africa. It should be noted that regional patterns of paleo-rainfall changes are in general difficult to simulate. In response to Last Glacial Maximum boundary conditions, different coupled general circulation models simulate widely varying responses in rainfall over Africa<sup>28</sup>.

## **Reporting summary**

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

# **Data availability**

The consensus sequences for this set of 198 mitogenomes have been deposited in the NCBI GenBank with accession numbers MK248274–MK248471. Requests for materials should in the first instance be addressed to V.M.H.

- Orizio, R. Lost White Tribes: the End of Privilege and the Last Colonials in Sri Lanka, Jamaica, Brazil, Haiti, Namibia, and Guadeloupe (Free, 2001).
- Heine, B. & Nurse, D. (eds) African Languages: an Introduction (Cambridge Univ. Press, 2000)
- Montinaro, F. et al. Complex ancient genetic structure and cultural transitions in southern African populations. Genetics 205, 303–316 (2017).
- 41. Guthrie, M. The Classification of the Bantu Languages (Oxford Univ. Press, 1948).
- Honken, H. & Heine, B. The Kx'a family: a new Khoisan genealogy. J. Asian Afr. Stud. 79, 5–36 (2010)
- Güldemann, T. & Elderkin, E. D. in Khoisan Languages and Linguistics: Proc. 1st International Symposium January 4–8, 2003, Riezlern/Kleinwalsertal (eds Brenzinger, M. & König. C.) 15–52 (Rüdiger Köppe. 2010).
- Stockton, R. The Herero genocide: Germany's first mass murder. All That's Interesting https://allthatsinteresting.com/herero-genocide (2017).
- Smith, A. B. Excavations at Kasteelberg and the Origins of the Khoekhoen in the Western Cape, South Africa (Archaeopress, 2006).
- Patterson, N. et al. Genetic structure of a unique admixed population: implications for medical research. Hum. Mol. Genet. 19, 411–419 (2010).
- 47. van der Ross, R. E. Up from Slavery: Slaves at the Cape: their Origins, Treatment and
- Contribution (Ampersand, 2005).

  48. McCrow, J. P. et al. Spectrum of mitochondrial genomic variation and associated clinical
- presentation of prostate cancer in South African men. Prostate 76, 349–358 (2016).
   Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27, 2987–2993 (2011).
- Weissensteiner, H. et al. HaploGrep 2: mitochondrial haplogroup classification in the era
  of high-throughput sequencing. Nucleic Acids Res. 44, W58–W63 (2016).
- van Oven, M. & Kayser, M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum. Mutat. 30, E386–E394 (2009).
- 52. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
- Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5, e9490 (2010).
- Bouckaert, R. et al. BEAST 2: a software platform for Bayesian evolutionary analysis. PLOS Comput. Biol. 10, e1003537 (2014).
- Kivisild, T. et al. The role of selection in the evolution of human mitochondrial genomes. Genetics 172, 373–387 (2006).
- Herrnstadt, C. et al. Reduced-median-network analysis of complete mitochondrial DNA coding-region sequences for the major African, Asian, and European haplogroups. Am. J. Hum. Genet. 70, 1152–1171 (2002).
- Schuster, S. C. et al. Complete Khoisan and Bantu genomes from southern Africa. Nature 463, 943–947 (2010).
- Soares, P. et al. Correcting for purifying selection: an improved human mitochondrial molecular clock. Am. J. Hum. Genet. 84, 740–759 (2009).
- Briggs, A. W. et al. Targeted retrieval and analysis of five Neandertal mtDNA genomes Science 325, 318–321 (2009).
- Green, R. E. et al. The Neandertal genome and ancient DNA authenticity. EMBO J. 28, 2494–2502 (2009).
- Prüfer, K. et al. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505, 43–49 (2014).
- Gandini, F. et al. Mapping human dispersals into the Horn of Africa from Arabian Ice Age refugia using mitogenomes. Sci. Rep. 6, 25472 (2016).
- Soares, P. et al. The expansion of mtDNA haplogroup L3 within and out of Africa. Mol. Biol. Evol. 29, 915–927 (2012).
- Friedrich, T., Timmermann, A., Tigchelaar, M., Elison Timm, O. & Ganopolski, A. Nonlinear climate sensitivity and its implications for future greenhouse warming. Sci. Adv. 2, e1501923 (2016)
- Stockhecke, M. et al. Millennial to orbital-scale variations of drought intensity in the Eastern Mediterranean. Quat. Sci. Rev. 133, 77–95 (2016).
- Laskar, J. et al. A long-term numerical solution for the insolation quantities of the Earth. Astron. Astrophys. 428, 261–285 (2004).
- Barbieri, C. et al. Unraveling the complex maternal history of Southern African Khoisan populations. Am. J. Phys. Anthropol. 153, 435–448 (2014).
- Barbieri, C., Butthof, A., Bostoen, K. & Pakendorf, B. Genetic perspectives on the origin of clicks in Bantu languages from southwestern Zambia. Eur. J. Hum. Genet. 21, 430–436 (2013).

- Barbieri, C. et al. Contrasting maternal and paternal histories in the linguistic context of Burkina Faso. Mol. Biol. Evol. 29, 1213–1223 (2012).
- Barbieri, C. et al. Migration and interaction in a contact zone: mtDNA variation among Bantu-speakers in Southern Africa. PLoS ONE 9, e99117 (2014).
- Batini, C. et al. Insights into the demographic history of African Pygmies from complete mitochondrial genomes. Mol. Biol. Evol. 28, 1099–1110 (2011).
- Eaaswarkhanth, M. et al. Traces of sub-Saharan and Middle Eastern lineages in Indian Muslim populations. Eur. J. Hum. Genet. 18, 354–363 (2010).
- Gonder, M. K., Mortensen, H. M., Reed, F. A., de Sousa, A. & Tishkoff, S. A. Whole-mtDNA genome sequence analysis of ancient African lineages. *Mol. Biol. Evol.* 24, 757–768 (2007)
- Horai, S., Hayasaka, K., Kondo, R., Tsugane, K. & Takahata, N. Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. Proc. Natl Acad. Sci. USA 92, 532–536 (1995).
- Ingman, M., Kaessmann, H., Pääbo, S. & Gyllensten, U. Mitochondrial genome variation and the origin of modern humans. *Nature* 408, 708–713 (2000).
- Just, R. S., Diegoli, T. M., Saunier, J. L., Irwin, J. A. & Parsons, T. J. Complete mitochondrial genome sequences for 265 African American and U.S. "Hispanic" individuals. Forensic Sci. Int. Genet. 2, e45–e48 (2008).
- Kujanová, M., Pereira, L., Fernandes, V., Pereira, J. B. & Cerný, V. Near eastern Neolithic genetic input in a small oasis of the Egyptian Western Desert. Am. J. Phys. Anthropol. 140, 336–346 (2009).
- Maca-Meyer, N., González, A. M., Larruga, J. M., Flores, C. & Cabrera, V. M. Major genomic mitochondrial lineages delineate early human expansions. BMC Genet. 2, 13 (2001).
- Macaulay, V. et al. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science 308, 1034–1036 (2005).
- Margaryan, A. et al. Eight millennia of matrilineal genetic continuity in the South Caucasus. Curr. Biol. 27, 2023–2028 (2017).
- Olivieri, A. et al. Mitogenome diversity in Sardinians: a genetic window onto an island's past. Mol. Biol. Evol. 34, 1230–1239 (2017).
- van der Walt, E. M. et al. Characterization of mtDNA variation in a cohort of South African paediatric patients with mitochondrial disease. Eur. J. Hum. Genet. 20, 650–656 (2012).
- Vyas, D. N. et al. Bayesian analyses of Yemeni mitochondrial genomes suggest multiple migration events with Africa and Western Eurasia. Am. J. Phys. Anthropol. 159, 382–393 (2016).

Acknowledgements We thank all of the study participants, as well as the many people who provided assistance during participant recruitment and recording, or provided critical historical, cultural and linguistic insights including; C. P. Bennett (https://evolvingpicture. com/), R. Wilkinson, J. Sinvula, H. Money, the late C. F. Heyns, R. H. Glashoff, D. de Swart, P. Fernandez, P. A. Venter, S. C. Schuster, M. P. Marx, the late S. M. Kooitije (39th leader of the ‡Aonin clan and chairperson of the Nama Traditional Leaders Association), A. A. Collins, B. Kaesie, J. Kavimbi, H. Mische, F. Nague, D. Nague, H. Oosthuizen, E. Oosthuizen, A. Oosthuysen, E. Oosthuysen, D. Roux, C. Swau and T. Tsebe. We acknowledge the late M. McFarlane, who identified Deception ridge and its importance in the evolution of the Makgadikgadi palaeo-lake, This work was supported by an Australian Research Council Discovery Project grant awarded to V.M.H. (DP170103071) and sampling contributed by the Cancer Association of South Africa to M.S.R.B. and V.M.H. A.T. and S.-S.L. received funding from the Institute for Basic Science (IBS) under IBS-R028-D1, V.M.H. is supported by the University of Sydney Foundation in a Petre Foundation chair position. Computational resources were provided by the Australian Government through the National Computational Infrastructure, the Sydney Informatics Research Hub at the University of Sydney (Artemis HPC) and by the Garvan Institute of Medical Research Data Intensive Computer Engineering team.

Author contributions V.M.H. designed the study. M.S.R.B., H.E.A.F. and V.M.H. obtained and maintain study approvals and permits, as well as community leadership support. M.S.R.B., D.C.P. and V.M.H. performed recruitments, consenting, sampling and processing. R.J.L., A.M.F.K. and D.C.P. performed pre-screening and mitogenome data generation. E.K.F.C. performed the bioinformatics and phylogenetic analyses. A.E.M. performed geographical interpretation. S.-S.L. and A.T. performed climatological model analyses and interpretation, with additional local climatology interpretation provided by H.R. V.M.H. led the interpretation of the multiple-discipline analyses, with contributions from all of the authors. E.K.F.C., B.F.B., S.-S.L., A.T. and V.M.H. generated and interpreted the figures. V.M.H., E.K.F.C. and A.T. wrote the manuscript with contributions from all of the authors.

Competing interests The authors declare no competing interests.

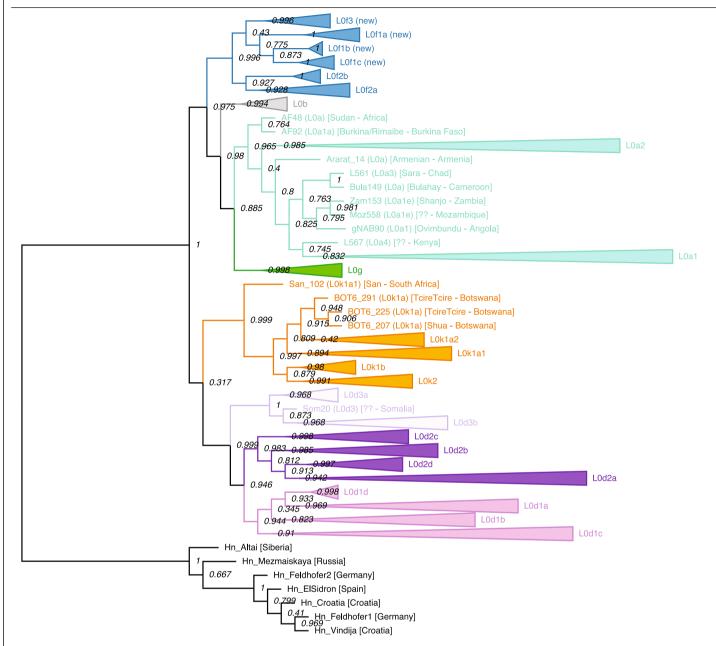
### Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1714-1.

Correspondence and requests for materials should be addressed to A.T. or V.M.H.

Peer review information Nature thanks Victor Brovkin, Rebecca Cann and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

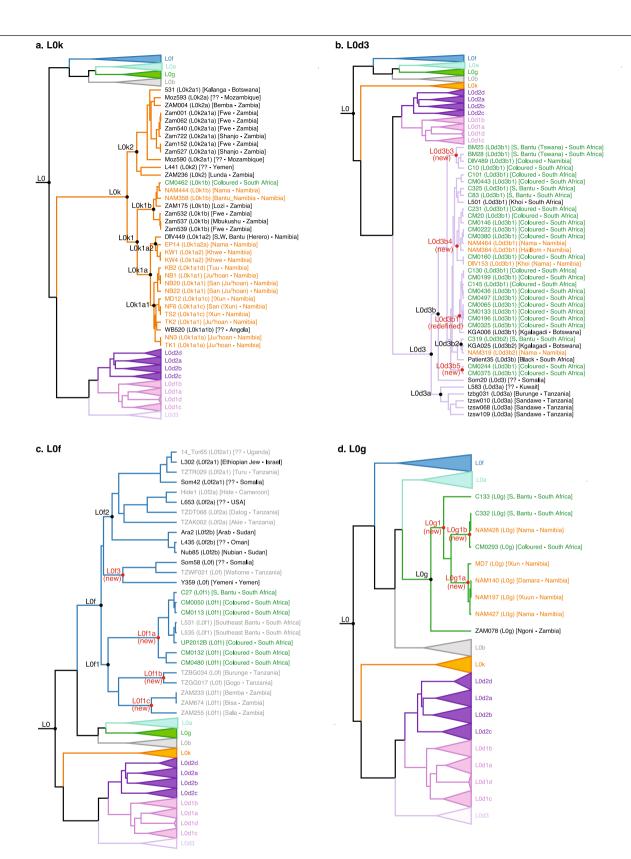
Reprints and permissions information is available at http://www.nature.com/reprints.



# $Extended\,Data\,Fig.\,1|\,Phylogenetic\,tree\,of\,all\,1, 217\,L0\,mitogenomes.$

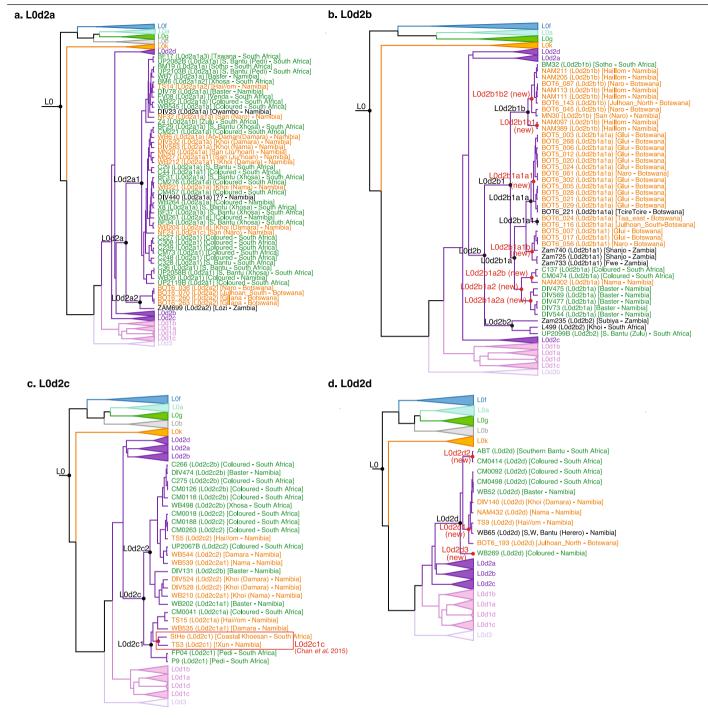
 $Phylogeny was inferred using FastTree \ v. 2.1.746, displayed using FigTree. Tips belonging to the same haplogroup are collapsed and coloured as in Fig. 2a. Local properties of the same haplogroup are collapsed and coloured as in Fig. 2a. Local properties of the same haplogroup are collapsed and coloured as in Fig. 2a. Local properties of the same haplogroup are collapsed and coloured as in Fig. 2a. Local properties of the same haplogroup are collapsed and coloured as in Fig. 2a. Local properties of the same haplogroup are collapsed and coloured as in Fig. 2a. Local properties of the same haplogroup are collapsed and coloured as in Fig. 2a. Local properties of the same haplogroup are collapsed and coloured as in Fig. 2a. Local properties of the same haplogroup are collapsed and coloured as in Fig. 2a. Local properties of the same haplogroup are collapsed and coloured as in Fig. 2a. Local properties of the same haplogroup are collapsed and coloured as in Fig. 2a. Local properties of the same haplogroup are collapsed and coloured as in Fig. 2a. Local properties of the same haplogroup are collapsed and coloured as in Fig. 2a. Local properties of the same haplogroup are collapsed and coloured as in Fig. 2a. Local properties of the same haplogroup are collapsed and coloured as in Fig. 2a. Local properties of the same haplogroup are collapsed and coloured as in Fig. 2a. Local properties of the same haplogroup are collapsed and coloured as in Fig. 2a. Local properties of the same haplogroup are collapsed as in Fig. 2a. Local properties of the same haplogroup are collapsed as in Fig. 2a. Local properties of the same haplogroup are collapsed as in Fig. 2a. Local properties of the same haplogroup are collapsed as in Fig. 2a. Local properties of the same haplogroup are collapsed and col$ 

support values for each node are indicated and branch lengths are proportional to the number of substitutions per site. The tree is rooted to the seven Neanderthal mitogenomes as indicated.



Extended Data Fig. 2 | Detailed phylogenetic branching of L0k, L0d3, L0f and L0g. a–d, Expanded sections of the phylogenetic tree depicted in Fig. 2a are shown, including 34 (out of a total of 113) L0k (a), all 40 L0d3 (b), all 27 L0f (c) and all 9 L0g (d) mitogenomes. Each mitogenome is represented as a tip and coloured based on their broad ethno-linguistic classification, if known. KhoeSan

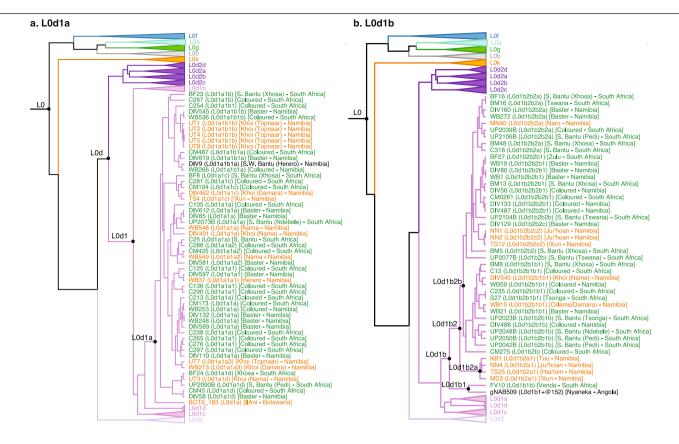
is shown in orange, non-KhoeSan in grey and Cape multi-ethnic (KhoeSan ancestral) in green. Publicly available mitogenomes for which we cannot be certain of their broad population identifier are labelled in black font. Proposed new sub-lineages for LOd3, LOf and LOg1 are indicated by red-coloured node labels and are further described in Supplementary Tables 7–9.

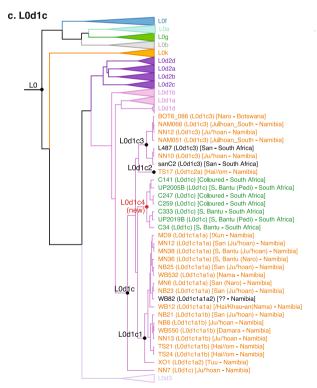


# $Extended\,Data\,Fig.\,3\,|\,Detailed\,phylogenetic\,branching\,of\,L0d2.$

 $\label{eq:acceleration} \textbf{a}, \textbf{c}, \textbf{d}, \text{Expanded branches of the phylogenetic tree depicted in Fig. 2a are shown, including 51 (out of a total of 118) L0d2a (\textbf{a}), 25 (out of 53) L0d2c (\textbf{c}) and all 11L0d2d (\textbf{d}) mitogenomes. \textbf{b}, For L0d2b, an additional BEAST analysis was performed using an alternate subset of 441 mitogenomes that included all 43 L0d2b samples, as opposed to the \textit{n} = 461 subset (Fig. 2a) that included $1.00 + 1.00 +$ 

only 13 L0d2b. The same model parameters were used for both data subsets. In all panels, each mitogenome is represented as a tip and coloured based on their broad ethno-linguistic classification, as in Extended Data Fig. 2. The previously defined L0d2c1c haplogroup, containing the coastal KhoeSan StHe skeleton and other newly proposed sub-lineages are indicated by red node labels (Supplementary Tables 4–6).

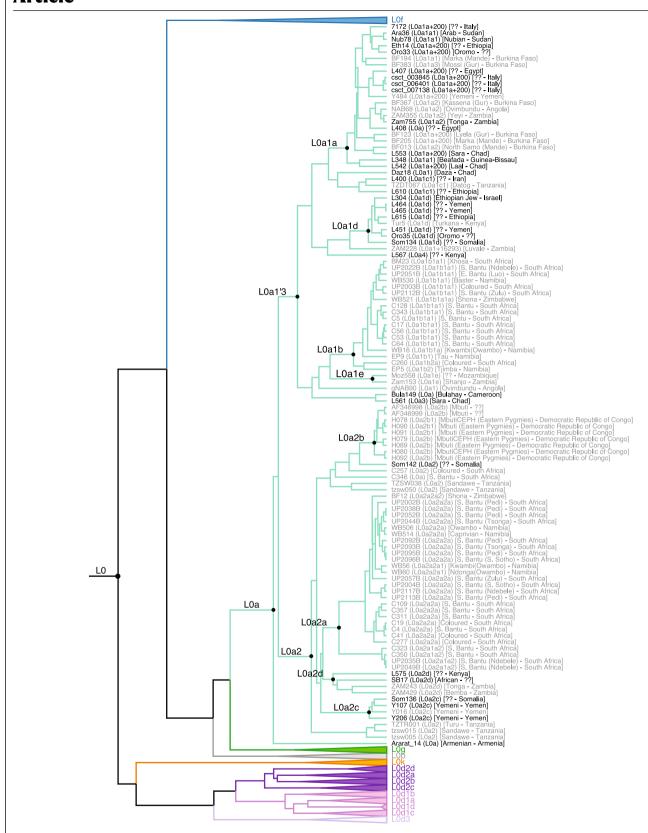




# $Extended\,Data\,Fig.\,4\,|\,Detailed\,phylogenetic\,branching\,of\,L0d1.$

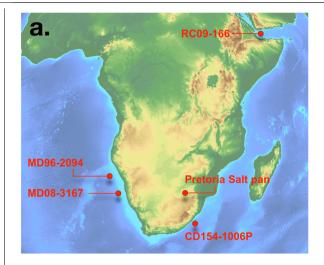
a-c, Expanded branches of the phylogenetic tree depicted in Fig.~2a are shown, including 54 (out of a total of 91) LOd1a (a), 45 (out of 174) LOd1b (b) and 33 (out of 174) LOd1b (b) and 33 (out of 174) LOd1b (c) and 33 (out of 174) LOd1b (d) and 34 (out of 174) LOd1b (d) an

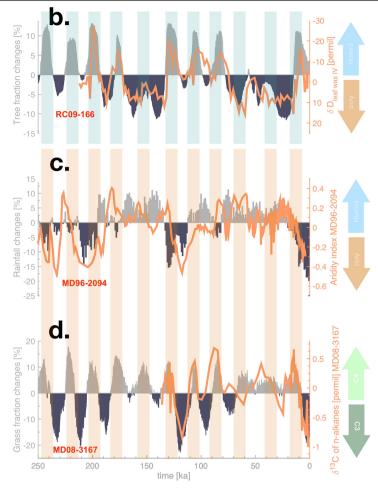
 $184) \, LOdic \, \textbf{(c)} \, mitogenomes. Each mitogenome is represented as tips and coloured based on their broad ethno-linguistic classification as in Extended Data Fig. 2.$ 



**Extended Data Fig. 5** | **Detailed phylogenetic branching of L0a.** The L0a branch of the phylogenetic tree displayed in Fig. 2a is shown, which includes a subset of 114 (out of a total of 294) L0a mitogenomes. Each mitogenome is

represented as tips and coloured based on their broad ethno-linguistic classification as in Extended Data Fig. 2.





#### $\label{lem:extended} \textbf{Extended Data Fig. 6} \ | \ \textbf{Comparison of the palaeo-data and palaeo-model.}$

 $\label{eq:andpalaeo-model} \textbf{a}. Locations of key sites that are used for the comparison of the palaeo-model and palaeo-data in this study are highlighted in red. The map was generated in Paraview v.5.6 (https://www.paraview.org/). \textbf{b}, Simulated tree fraction (%) at Horn of Africa (land grid points nearest to RC09-166) (grey, dark-blue bars) and stable hydrogen isotopic composition of leaf wax, corrected for ice volume contributions from the Gulf of Aden marine sediment core RC09-166 of (orange), indicating changes in hydroclimate. \textbf{c}, Relative precipitation changes (%)$ 

simulated by LOVECLIM transient model (all forcings) for  $11^{\circ}$  E,  $19^{\circ}$  S (grey, dark-blue bars) and grain-size aridity index reconstructed from sediment core MD96-2094 corange). **d**, Grass fraction changes simulated by LOVECLIM transient model (all forcings) at  $11^{\circ}$  E, 14– $17^{\circ}$  S (grey, dark-blue bars) and reconstructed  $\delta^{13}$ C changes of n-alkanes (orange) (South Atlantic sediment core MD08-3167) indicative of abundance of  $C_3$  and  $C_4$  plants in the Namibian desert and further inland  $S_3$ .

# Extended Data Table 1 | LO mitogenomes included in this study

	Number of L0
	Mitogenomes
Barbieri <i>et al</i> ., <b>AJHG</b> 2013 <sup>27</sup>	485
Barbieri <i>et al</i> ., <b>AJPA</b> 2014 <sup>65</sup>	26
Barbieri <i>et al</i> ., <b>EJHG</b> 2013 <sup>66</sup>	33
Barbieri <i>et al</i> ., <b>MBE</b> 2012 <sup>67</sup>	10
Barbieri <i>et al</i> ., <b>PlosOne</b> 2014 <sup>68</sup>	115
Batini <i>et al</i> ., <b>MBE</b> 2011 <sup>69</sup>	9
Behar <i>et al</i> ., <b>AJHG</b> 2008 <sup>1</sup>	64
Chan <i>et al</i> ., <b>Plos One</b> 2015 <sup>6</sup>	77 *
Eaaswarkhanth et al., <b>EJHG</b> 2009 <sup>70</sup>	3
Gonder <i>et al.</i> , <b>MBE</b> 2007 <sup>71</sup>	27
Herrnstadt <i>et al</i> ., <b>AJHM</b> 2002 <sup>55</sup>	2 †
Horai <i>et al</i> ., <b>PNAS</b> 1995 <sup>72</sup>	1‡
Ingman <i>et al</i> ., <b>Nature</b> 2000 <sup>73</sup>	5
Just <i>et al</i> ., <b>FSIG</b> 2008 <sup>74</sup>	8
Kivisild <i>et al</i> ., <b>Genetics</b> 2006 <sup>54</sup>	25 †
Kujanova <i>et al</i> ., <b>AJPA</b> 2009 <sup>75</sup>	4 §
Maca-Meyer et al., BMC Genet 200	)1 <sup>76</sup> 1
Macaulay et al., Science 2005 <sup>77</sup>	1
Margaryan <i>et al</i> ., <b>Curr Biol</b> 2017 <sup>78</sup>	1
McCrow et al., Prostate 2016 <sup>47</sup>	37 *
Morris <i>et al.</i> , <b>GBE</b> 2014 <sup>15</sup>	1 *
Olivieri <i>et al</i> ., <b>MBE</b> 2017 <sup>79</sup>	4
Rito et al., Plos One 2013 <sup>3</sup>	42
Schuster et al., Nature 2010 <sup>56</sup>	6 *
van der Walt <i>et al</i> ., <b>EJHG</b> 2012 <sup>80</sup>	20
Vyas <i>et al</i> ., <b>AJPA</b> 2016 <sup>81</sup>	12
Total Public	1,019
This Study	198
Total L0 Mitogenomes	1,217

Numbers of mitogenomes taken from previously published  $^{67\text{-}83}$  studies.

<sup>\*</sup>Previously published data by our group with verified population metadata.

<sup>&</sup>lt;sup>†</sup>Mitochondrial DNA sequences of the coding-region only.

<sup>\*</sup>Sequence has non-canonical start position corresponding to position 577 of rCRS.

<sup>§</sup>Coriell cell lines.

# Extended Data Table 2 | KhoeSan population identifiers used in this study

Broad identifier	Geographic identifier	Broad language group	Ethno-linguistic identifiers
KhoeSan	Kalahari KhoeSan	Kw'a	Ju/'hoansi; !Xun; !Xuun; Hoan;
		Tuu	/Hoa; Taa
		Khoe-Kwadi (central)	Naro;   Ani; Buga; G  ana; G ui; Khwe; Shua; Tshwa;   Xokhoe
	West-coastal KhoeSan	Khoe-Kwadi (Nama)	Nama; Damara; #Anonin; Ao-Daman; Hai  om; Orlams
KhoeSan ancestral	Cape KhoeSan	Afrikaans	Baster; Coloured
	East-Coastal KhoeSan	Southern Bantu	Ndebele; Pedi; Tswana; Tsonga; Venda; Xhosa; Zulu



Corresponding author(s):	Vanessa M. Hayes
Last updated by author(s):	2019/09/04

# **Reporting Summary**

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

_					
C 1	- ~	+1	st	117	$\sim$ C
- N I	_		<b>^1</b>	-11	_

FUI	an statistical analyses, commit that the following items are present in the rigure regend, table regend, main text, or Methods Section.
n/a	Confirmed
	The exact sample size $(n)$ for each experimental group/condition, given as a discrete number and unit of measurement
	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
	The statistical test(s) used AND whether they are one- or two-sided Only common tests should be described solely by name; describe more complex techniques in the Methods section.
	A description of all covariates tested
	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
$\boxtimes$	For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i> ) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted <i>Give P values as exact values whenever suitable.</i>
	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
$\boxtimes$	Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i> ), indicating how they were calculated
	Our was collection on statistics for biologists contains articles on many of the points above

Our web collection on <u>statistics for biologists</u> contains articles on many of the points above.

## Software and code

Policy information about availability of computer code

Data collection

The nucleotide sequences of 1,019 LO mitogenomes were downloaded from NCBI

Data analysis

All software used is described and referenced in the METHODS section. No new software algorithms or code was established. These include:

Mitogenomic and phylogenetic analyses:

Consensus mitogenome sequences were derived by first identifying variants relative to rCRS using samtools (v1.3.1) mpileup (with parameters -d 10000 -L 1000 -Q 7 -h 50 -o 10 -e 17 -m 4) and beftools (v1.3.1). Variant data was converted to fasta format using samtools' vefutils.pl vcf2fq. Mitochondrial haplogroups were evaluated (called) using HaploGrep2 (v2.1.13) based on PhyloTree Build 17. Multiple sequence alignment (for phylogenetic inference) was performed using FastTree v2.1.7 (SSE3) using the generalised time reversible (-gtr) and discrete gamma model with 20 rate categories (-gamma). Bayesian phylogenetic inferences and divergence times were calculated using BEAST2 v2.4.2 with BEAGLE 2.0. Multiple sequence alignment of the subset of 461 AMH and seven Neanderthal mitogenomes was converted to NEXUS format using the convert function of seqmagick v0.6.1 (https://fhcrc.github.io/seqmagick) with parameter --alphabet dna-ambiguous. This provided the input to BEAST2. Specifically, BEAUTi v2.4.2 was used to set up the phylogenetic model, assuming: (i) the Gamma Site Model with 6 gamma categories and no invariant sites; (ii) the generalized time reversible substitution model; (iii) a strict constant clock model with a normal prior of with  $\mu$  = 1.665 x 10-8 and  $\sigma$  = 1.479 x 10-9 based on Soares et al. 200960; and (iv) a Coalescent Constant Population. Five BEAST replicates were performed, each with 100 million MCMC iterations, sampling every 10,000. Tracer v1.6 was used to evaluate BEAST trace files. Replicates were combined using LogCombiner v2.4.2. Sampled trees from BEAST were summarized into a single Maximum Clade Credibility target tree using TreeAnnotator v2.4.2 for each of the five replicates, discarding the first 10% as burn-ins. FigTree v1.4.2+ was used to visualize all resulting trees.

Bayesian Skyline Plot (BSP) analysis: For each haplogroup of interest (e.g. LOa, LOd1'2, and LOk), a nexus file was derived using SeqMagic v0.6.1 as described above. BSP analyses were performed using BEAST2, using BEAUTi 2 for model setup as before, with the following key differences: (i) the gamma shape of the Gamma Site Model was estimated with an exponential prior with mean = 1.0 and offset = 0.0; (ii) the molecular clock was fixed (not estimated) at 1.665 x 10-8 based on Sores et al.60; and (iii) the phylogenetic tree prior was set to

Coalescent Bayesian Skyline, assuming 20 intervals between the root of the tree and the present time. Tracer v1.6 was used to reconstruct the Bayesian Skyline from the sampled trees for each analysis, using a stepwise constant variant and the lower 95% highest posterior density of the root height as the maximum time.

Climate model: LOVECLIM earth system model

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

#### Data

Randomization

Blinding

Policy information about <u>availability of data</u>

All manuscripts must include a <u>data availability statement</u>. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The consensus sequences for this set of 198 mitogenomes have been deposited to NCBI with Accession Numbers MK248274-MK248471.

Field-spe	ecific reporting
Please select the o	ne below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.
Life sciences	Behavioural & social sciences Ecological, evolutionary & environmental sciences
For a reference copy of t	the document with all sections, see <a href="mailto:nature.com/documents/nr-reporting-summary-flat.pdf">nature.com/documents/nr-reporting-summary-flat.pdf</a>
Life scier	nces study design
All studies must dis	sclose on these points even when the disclosure is negative.
Sample size	No sample size calculation was performed as all mitogenomes representing a rarer LO haplogroup were included. Pre-screening was performed on roughly 500 population relevant samples, identifying 198 of interest to undergo complete mitogenome sequencing and inclusion in this study. All relevant published data was also downloaded for a total study of 1,217. Sample sizes are therefore ALL inclusive (no rare LO lineage and mitogenome was excluded) and therefore sufficient.
Data exclusions	All 1,217 mitogenomes were used to determine haplogroup frequencies and reconstruct geographic dispersals. A focused subset of 461
Data exclusions	mitogenomes, including all rare lineages, were used to establish within LO coalescence times. The smaller subset was necessary due to the computational burden of the BEAST analysis.
Replication	LO haplogroup pre-screening (2,673 bp region) using Sanger sequencing for the 198 mitogenomes that underwent Ion Torrent deep complete
	mitogenome sequencing, providing an internal experimental validation. All samples were validated.

# Reporting for specific materials, systems and methods

self-identification (ethnic classifications) of participating subjects and country of origin.

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Not relevant to this study as grouping (haplogroups) were identified via mitogenomic data. KhoeSan geographic classification were based on

Investigators were blinded to the subject identifiers during analysis, as were the climate physicists blinded to the hypotheses and distributions

Materials & experimental systems		Methods	
n/a	Involved in the study	n/a	Involved in the study
$\boxtimes$	Antibodies	$\boxtimes$	ChIP-seq
$\boxtimes$	Eukaryotic cell lines	$\boxtimes$	Flow cytometry
	Nalaeontology	$\boxtimes$	MRI-based neuroimaging
$\boxtimes$	Animals and other organisms		
	Human research participants		
$\boxtimes$	Clinical data		
	•		

associated with the mitogenomic data.

## Palaeontology

Specimen provenance

NA - no specimens were collected for the purpose of this study - only published data used.

Specimen deposition

NA - the study involved published data

Dating methods

No new specimens or dates are provided or used.

| Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

## Human research participants

Policy information about studies involving human research participants

Population characteristics

Southern Africans from Namibia and South Africa representing and self-identifying as ancestrally from a KhoeSan or KhoeSan ancestral population identifiers (as outlined within the METHODS) were recruited. There was no gender bias and all participants were greater than 18 years of age as per ethical requirements.

Recruitment

Participants were recruited based on their self-identified ethnicity. Recruitment was led by local researchers and coauthors (MSRB, DCP or VMH in South Africa) or in Namibia (HATF, as well as VMH). The study was explained to the participants and communities and in some communities, especially contemporary Namibian Kalahari KhoeSan populations, this took place over an extensive period, with regular engagement with the communities by VMH over a 10 year period. In these communities recruitment was a decision made by the entire community. All recruiters are familiar with local languages and cultures. There was no other biases that would impact this study. All possible populations that may carry a L0 mitogenome and live within the borders of Namibia and South Africa and were willing to participate, were included.

Ethics oversight

The study was reviewed and approved by the Ministry of Health and Social Services (MoHSS) in Namibia (#17-3-3 2008, 2014 and 2019), with additional local approvals from community leaders, the University of Pretoria Human Research Ethics Committee (HREC #43/2010 and HREC #280/2017), including US Federal-wide assurance (FWA00002567 and IRB00002235 IORG0001762), as well as the South African National Blood Service (SANBS) HREC (HREC #2012/11). Isolated DNA was shipped under the Republic of South Africa Department of Health Export Permit (#J1/2/4/2), in accordance with the National Health Act 2003, to the Garvan Institute of Medical Research in Australia. Mitogenome sequencing was performed in accordance with site-specific approval granted by St Vincent's Hospital HREC in Australia (SVH 15/227).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

# Quantifying the dynamics of failure across science, startups and security

https://doi.org/10.1038/s41586-019-1725-y

Yian Yin<sup>1,2,3</sup>, Yang Wang<sup>1,2,4</sup>, James A. Evans<sup>5,6</sup> & Dashun Wang<sup>1,2,3,4</sup>\*

Received: 15 February 2019

Accepted: 27 September 2019

Published online: 30 October 2019

Human achievements are often preceded by repeated attempts that fail, but little is known about the mechanisms that govern the dynamics of failure. Here, building on previous research relating to innovation<sup>1-7</sup>, human dynamics<sup>8-11</sup> and learning<sup>12-17</sup>, we develop a simple one-parameter model that mimics how successful future attempts build on past efforts. Solving this model analytically suggests that a phase transition separates the dynamics of failure into regions of progression or stagnation and predicts that, near the critical threshold, agents who share similar characteristics and learning strategies may experience fundamentally different outcomes following failures. Above the critical point, agents exploit incremental refinements to systematically advance towards success, whereas below it, they explore disjoint opportunities without a pattern of improvement. The model makes several empirically testable predictions, demonstrating that those who eventually succeed and those who do not may initially appear similar, but can be characterized by fundamentally distinct failure dynamics in terms of the efficiency and quality associated with each subsequent attempt. We collected large-scale data from three disparate domains and traced repeated attempts by investigators to obtain National Institutes of Health (NIH) grants to fund their research, innovators to successfully exit their startup ventures, and terrorist organizations to claim casualties in violent attacks. We find broadly consistent empirical support across all three domains, which systematically verifies each prediction of our model. Together, our findings unveil detectable yet previously unknown early signals that enable us to identify failure dynamics that will lead to ultimate success or failure. Given the ubiquitous nature of failure and the paucity of quantitative approaches to understand it, these results represent an initial step towards the deeper understanding of the complex dynamics underlying failure.

To understand the dynamics of failure, we collected three large-scale datasets (Supplementary Information 1). The first dataset ( $D_i$ ) contains all R01 grant applications submitted to the NIH (776,721 applications by 139,091 investigators, 1985–2015; Supplementary Information 1.1). For each grant application, we obtained ground-truth information on whether or not it was funded, allowing us to reconstruct individual application histories and their repeated attempts to obtain funding. Our second dataset  $(D_2)$  traces start-up investment records from VentureXpert18 (58,111 startup companies involving 253,579 innovators, 1970–2016; Supplementary Information 1.2). Tracing every startup in which venture capital firms invested,  $D_2$  allows us to reconstruct individual career histories counting successive ventures in which they were involved. Here we follow previous studies in the entrepreneurship literature<sup>19</sup>, and classify successful ventures as those that achieved initial public offering (IPO) or high-value mergers and acquisitions, and correspondingly failed attempts as those that failed to obtain

such an exit within five years after their first investment by venture capital firms. Going beyond traditional innovation domains, we collected our third dataset ( $D_3$ ) from the Global Terrorism Database<sup>20</sup> (170,350 terrorist attacks by 3,178 terrorist organizations, 1970–2017; Supplementary Information 1.3). For each organization we trace their attack histories<sup>21,22</sup>, and classify success as fatal attacks that killed at least one person, and correspondingly failure as those that failed to claim casualties.

#### Mechanisms of chance and learning

Chance and learning  $^{13,16}$  are two primary mechanisms that explain how failures may lead to success. If each attempt has a certain likelihood of success, the probability that multiple attempts all lead to failure decreases exponentially with each trial. The chance model therefore emphasizes the role of luck, suggesting that success eventually arises

<sup>1</sup>Center for Science of Science and Innovation, Northwestern University, Evanston, IL, USA. <sup>2</sup>Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL, USA. <sup>3</sup>McCormick School of Engineering, Northwestern University, Evanston, IL, USA. <sup>4</sup>Kellogg School of Management, Northwestern University, Evanston, IL, USA. <sup>5</sup>Department of Sociology, University of Chicago, Chicago, IL, USA. <sup>6</sup>Santa Fe Institute, Santa Fe, NM, USA. \*e-mail: dashun.wang@northwestern.edu

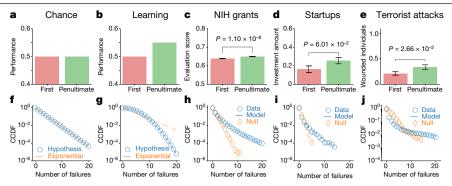


Fig. 1 | Mechanisms of chance and learning. a-i, We compare theoretical predictions and empirical measurements for performance changes (a-e) as well as the length distribution of failure streaks (f-j). a, f, The chance model predicts no performance change (a) with a failure streak length that follows an exponential distribution (f). b, g, The learning hypothesis predicts improved performance (b) with failure streaks that are shorter than expected by the chance model, corresponding to a faster-than-exponential distribution (g). Both hypotheses are contested by empirical patterns observed across the three datasets. To ensure that performance metrics are comparable across data and models, we standardized performance measures according to their underlying distribution (Supplementary Information 5.1). c-e, We find that failures in real data are associated with improved performance between the first and

penultimate attempt. Two-sided Welch's t-test; data are mean ± s.e.m. **c**, n = 4,872 (first), 5,966 (penultimate). **d**, n = 579 (first), 548 (penultimate).  $\mathbf{e}$ , n = 231 (first), 230 (penultimate).  $\mathbf{h} - \mathbf{j}$ , At the same time, however, failure streaks are characterized by a fat-tailed length distribution, indicating that failure streaks in real data are longer than expected by chance. For clarity, here we show results for failure streaks for which the length is less than 21 (Supplementary Information 5.2). We further construct a randomized sequence of successes and failures by assigning each attempt to agents at random (Supplementary Information 5.2). We find that failure streak length in the randomized sequence follows an exponential-like distribution, showing clear deviations from the data.

from an accumulation of independent trials. To test this, we compared the performance of the first and penultimate attempt within failure streaks (Supplementary Information 5.1), measured by NIH percentile score for a grant application  $(D_1)$ , investment size by venture capital firms to a company  $(D_2)$  and number of wounded individuals by an attack  $(D_3)$ . We find that across all three datasets, the penultimate attempt shows systematically better performance than the initial attempt (Fig. 1c-e). These results reject that success is simply driven by chance (Fig. 1a) but lend support to the learning mechanism (Fig. 1b), which suggests that failure may teach valuable lessons that are difficult to learn otherwise<sup>12,13,16</sup>. As such, learning reduces the number of failures required to achieve success, and predicts that failure streaks should follow a narrower length distribution (Fig. 1g) than the exponential distribution predicted by chance (Fig. 1f). However, across all three domains, the length of failure streaks follows a fat-tailed distribution (Fig. 1h-j, Supplementary Information 5.2), indicating that despite improvements in performance, failures are characterized by longer-than-expected streaks before the onset of success. Together, these observations demonstrate that neither chance nor learning alone can explain the empirical patterns that underlie failures, suggesting that more complex dynamics may be at work.

#### Modelling dynamics of failure

Here we explore the interplay between chance and learning by developing a simple one-parameter model that mimics how future attempts build on previous failures (Fig. 2a, b, Supplementary Information 3.1). We consider that each attempt consists of many independent, unweighted components, with each component i being characterized by an evaluation score xi (Fig. 2a). For example, components for the submission of an NIH proposal include constructing a biosketch, assembling a budget, writing a data management plan, adding preliminary data and outlining broader impacts. We also note that granting agencies often provide rubrics to grade proposals on specific components.

To formulate a new attempt, one goes through each component, and decides to either create a new version (with probability p) or reuse the best version  $x^*$  among the previous k attempts (with probability 1-p) (Fig. 2b). A new version is assigned a score drawn randomly from a uniform distribution U[0,1], approximating the percentile of score distributions real systems follow. The decision to create a new version

is often not random, but driven by the quality of previous versions. Indeed, given the best version  $x^*$ ,  $1-x^*$  captures the potential to improve it<sup>16</sup>. The higher this potential, the more likely one may create a new version, prompting us to consider a simple relationship,  $p = (1-x^*)^{\alpha}$ . with  $\alpha > 0$  (Methods, Supplementary Information 3.6). Creating a new version takes one unit of time with no certainty that its score will be higher or lower than the previous one. By contrast, reusing the best version from the past saves time, and allows the component to retain its best score x\*.

Here we explore a single parameter k for our model, measuring the number of previous attempts one considers when formulating a new one (Fig. 2b). Mathematically the dynamical process can be described as: with probability  $p, x_n \sim U[0, 1]$  or  $x_n = x_n^*$  otherwise (with probability (1-p) where  $x_n^* = \max\{x_{n-k}, \dots, x_{n-1}\}$ . We quantify the dynamics of the model by calculating the quality of the *n*th attempt,  $\langle x_n \rangle$ , which measures the average score of all components, and the efficiency after that attempt,  $\langle t_n \rangle$ , which captures the expected proportion of components updated in new versions. Let us first consider the two extreme cases. In the first case, k = 0 means that each attempt is independent from previous attempts (Supplementary Information 3.2). Here our model recovers the chance model, predicting that as *n* increases, both  $\langle x_n \rangle$  and  $\langle t_n \rangle$  remain constant (Extended Data Fig. 1a, d). That is, without considering past experience, failure does not lead to quality improvement. Nor is it more efficient to try again.

The other extreme  $(k \to \infty)$  considers all past attempts. The model predicts a temporal scaling in failure dynamics (Supplementary Information 3.3). That is, the time it takes to formulate a new attempt decays with *n*, asymptotically following a power law (Extended Data Fig. 1e):

$$T_n \equiv \langle t_n \rangle / \langle t_1 \rangle \sim n^{-\gamma} \tag{1}$$

where  $\gamma = \gamma_{\infty} = \alpha/(\alpha + 1)$  falls between 0 and 1 and '~' indicates 'asympototically proportional to'. Besides increased efficiency, new attempts also improve in quality, as the average potential for improvement decays according to  $\langle 1-x_n \rangle \sim n^{-\eta_{\infty}}$  where  $\eta_{\infty} = \min\{\gamma_{\infty}, 1-\gamma_{\infty}\}$  (Extended Data Fig. 1b). Here the model recovers the canonical result from the learning literature 12,15,23-25, commonly known as Wright's law 26. This is because, as experience accumulates, high-quality versions are preferentially retained, whereas their lower-quality counterparts are more likely to receive updates. As fresh attempts improve in quality (Extended Data

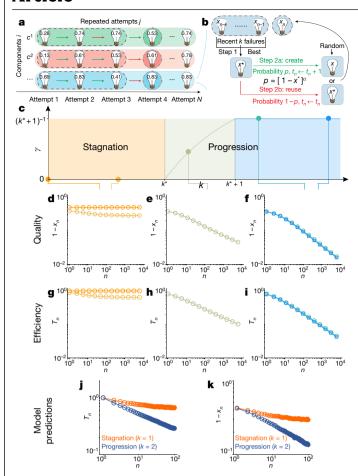


Fig. 2 | The k model. a, We treat each attempt as a combination of many independent components ( $c^i$ ). For attempt j, each component i is characterized by an evaluation score  $x_i^i$ , which falls between 0 and 1. The score for a new version is often unknown until attempted, hence a new version is assigned a score, drawn randomly from the range 0-1. **b**, To formulate a new attempt, one can either create a new version (with probability p, green arrow) or reuse an existing version by choosing the best one among past versions  $x^*$  (with probability 1-p, red arrow).  $P(x \ge x^*) = 1-x^*$  captures the potential to improve on prior versions, prompting us to assume  $p = (1 - x^*)^{\alpha}$  where  $\alpha > 0$  characterizes the propensity of an agent to create new versions given the quality of existing ones.  $\mathbf{c}$ , The analytical solution of the model reveals that the system is separated into three regimes by two critical points  $k^*$  and  $k^* + 1$ . The solid line shows the extended solution space of our analytical results. d-i, Simulation results from the model ( $\alpha = 0.6$ ) for quality (**d**-**f**) and efficiency (**g**-**i**) trajectories for different k parameters, showing distinct dynamical behaviour in different regimes. All results are based on simulations averaged over 104 times. j, k, A phase transition around  $k^*$  predicts the coexistence of two groups that fall in the stagnation and progression regimes, respectively.

Fig. 1b), they reduce the need to start anew, thus increasing the efficiency of future attempts (Extended Data Fig. 1e).

These two limiting cases (Extended Data Fig. 1c, f) might lead one to suspect a gradual emergence of scaling behaviour as we learn from more failures. By contrast, as we increase parameter k, the scaling exponent  $\gamma$  follows a discontinuous pattern (Fig. 2c, Supplementary Information 3.4) and only varies within a narrow interval of  $\lfloor k^* \rfloor < k < \lceil k^* \rceil + 1$  where  $k^* \equiv 1/\alpha$ . Indeed, when k is small ( $k < k^*$ ), the system converges back to the same asymptotic behaviour as k = 0 (Fig. 2c, d, g). In this region, k is not large enough to retain a good version once it appears. As a result, while performance might improve slightly in the first few attempts, it quickly saturates. In this region, agents reject previous attempts and flail around for new versions, not processing enough feedback to initiate a pattern of intelligent improvement, prompting

us to call it the stagnation region. Once k passes the critical threshold  $k^*$ , however, scaling behaviour emerges (Fig. 2c, e, h), indicating that the system enters a region of progression, in which failures lead to continuous improvement in both quality and efficiency. Nevertheless, with a single additional experience considered, the system quickly hits the second critical point  $k^*+1$ , beyond which the scaling exponent p becomes independent of k (Fig. 2c, f, i). This means that once  $\lfloor k^* \rfloor + 1$  number of previous failures is considered, the system is characterized by the same dynamical behaviour as  $k \to \infty$ , indicating that  $\lfloor k^* \rfloor + 1$  attempts are sufficient to recover the same rate of improvement as considering every failure from the past.

Importantly, the two critical points in our model can be mapped to phase transitions within a canonical ensemble consisting of three energy levels (Extended Data Fig. 1g–j, Methods, Supplementary Information 3.5). Phase transitions indicate that small variations at the microscopic level may lead to fundamentally different macroscopic behaviours. For example, two individuals near the critical point may initially appear identical in their learning strategy or other characteristics, but depending on which region they inhabit, their outcomes following failures could differ considerably (Fig. 2j, k). In the progression region  $(k > k^*)$ , agents exploit rapid refinements to improve through past feedback. By contrast, those in the stagnation region  $(k < k^*)$  do not seem to profit from failure, as their efforts stall in efficiency and saturate in quality. As such, the phase transitions uncovered in our simple model make four distinct predictions, which we now test directly in the contexts of science, entrepreneurship and security.

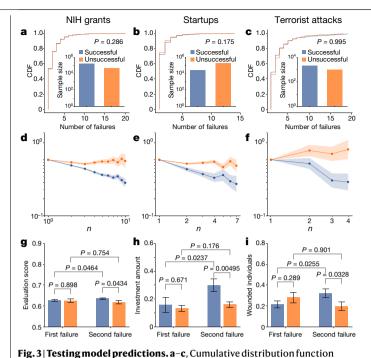
#### **Testing model predictions**

#### Not all failures lead to success

Although we tend to focus on examples that eventually succeeded following failures, the stagnation region predicts that there exists a non-negligible fraction of cases that do not succeed following failures. We measure the number of failed cases that did not achieve eventual success in our three datasets, finding not only that members of the unsuccessful group exist, but also that the size of the unsuccessful group is of a similar order of magnitude as the successful group (Fig. 3a-c). Notably, the number of consecutive failures before the last attempt for the unsuccessful group follows a statistically similar distribution from those that lead to success (Fig. 3a-c), suggesting that people who ultimately succeeded did not try more or less than their unsuccessful counterparts.

#### Early signals for ultimate success or failure

Our model predicts that the successful group is characterized by powerlaw temporal scaling (Eq. (1)), which is absent for the unsuccessful group (Fig. 2j), predicting that the two groups may follow fundamentally different failure dynamics that are distinguishable at an early stage. To test this prediction, we measure the average inter-event time between two failures  $T_n$  as a function of the number of failures (Supplementary Information 5.3). Figure 3d-f shows three important observations. First, for the successful group,  $T_n$  decays with n across all three domains, approximately following a power law, as captured by Eq. (1) (Extended Data Fig. 2, Supplementary Information 5.3, Supplementary Table 4). The scaling exponents are within a similar range as those reported in learning curves<sup>15</sup>, further supporting the validity of power-law scaling. Although the three datasets are among the largest in their respective domains, agents with a large number of failures are exceedingly rare, limiting the range of n that can be measured empirically. We therefore test whether alternative functions may offer a better fit, finding a power law to be the consistently preferred choice (Supplementary Information 6.2). Second, we found that temporal scaling disappears when we measure the same quantity for the unsuccessful group (Fig. 3d-f), consistent with predictions about the stagnation region. Third, the two groups show distinguishable failure dynamics as early as n = 2,



(CDF) of the number of consecutive failures before the last attempt for successful and unsuccessful groups. To eliminate the possibility that agents were simply in the process of formulating their next attempt, we focus on cases for which it has been at least five years since their last failure. In each of our three datasets, the two distributions are statistically indistinguishable (Kolmogorov-Smirnov test for samples with at least one failures). For clarity, here we show results for less than 21 failures (Supplementary Information 5.2). Inset, the sample size of successful and unsuccessful groups, showing their size is of a similar order of magnitude. **d-f**, Early temporal signals separate successful and unsuccessful groups.  $\mathbf{d}$ , n = 43,705 (successful), 15,132 (unsuccessful).  $\mathbf{e}$ , n = 2,455 (successful), 16,656 (unsuccessful).  $\mathbf{f}$ , n = 446(successful), 321 (unsuccessful). For each group, we measure the average interevent time between two failures  $T_n = t_n/t_1$  as a function of the number of attempts. Dots and shaded areas are mean ± s.e.m. measured from data (Supplementary Information 5.3). All successful groups manifest power-law scaling  $T_n \sim n^{-\gamma}$  (Extended Data Fig. 2). The two groups show distinguishable temporal dynamics for n = 2. Two-sided Welch's t-test;  $P = 3.02 \times 10^{-4}$ ,  $7.18 \times 10^{-3}$ , 9.42 × 10<sup>-2</sup> for comparisons of successful and unsuccessful groups in d, e, f respectively. This temporal scaling is absent for unsuccessful groups.  $\textbf{g-i}, Performance \, at \, first \, at tempt \, appears \, in distinguishable \, between \,$ successful and unsuccessful groups that experienced a large number of consecutive failures before the last attempt (at least 5 for  $D_1$ , 3 for  $D_2$  and 2 for  $D_3$ , two-sided Welch's t-test), but becomes distinguishable at the second attempt (two-sided Welch's t-test). Whereas performance improves for the successful group (one-sided Welch's t-test), this improvement is absent for the unsuccessful group (one-sided Welch's *t*-test). Data are mean  $\pm$  s.e.m. **g**, n = 628, 145, 571, 123 (from left to right). **h**, *n* = 248, 1, 332, 237, 1, 312 (from left to right). i, n = 231, 173, 229, 174 (from left to right).

suggesting noteworthy early signals that separate those who eventually succeed from those who do not.

Observations uncovered in Fig. 3d-f are notable for two main reasons. First, failures captured by the three datasets differ widely in their scope, scale, definition and temporal resolution, yet despite these differences, they are characterized by remarkably similar dynamical patterns predicted by our simple model. Second, although one might expect that the last attempt was crucial in separating the two groups, as the model predicts, successful and unsuccessful groups each follow their respective, highly predictable patterns, which are distinguishable long before the eventual outcome becomes apparent. Indeed, we use  $D_1$  to set up a prediction task (Extended Data Fig. 3, Methods, Supplementary Information 6.1) to predict ultimate success or failure using only temporal features, which yielded substantial predictive power. To test whether the observed patterns in Fig. 3d-f may simply reflect preexisting population differences, we take agents who experienced a large number of failures, and measure performance from their first attempt. We find that for all three domains, the two populations were statistically indistinguishable in their initial performance (Fig. 3g-i), which leads us to the next prediction.

#### Diverging patterns of performance improvement

Although the two groups may have begun with similar performances, the model predicts that they may experience different performance gains through failures (Fig. 2k). We compared performance at first and second attempts, finding significant improvement for the successful group (Fig. 3g-i), which is absent for the unsuccessful group. We further repeated our measurements by comparing the first and penultimate or halfway attempt, arriving at the same conclusion (Extended Data Fig. 9j-o, Supplementary Information 7.3). This prediction explains the patterns that were observed in Fig. 1c-e, which leads us to the second puzzle described in Fig. 1h-j: if performance improves, why are failure streaks longer than we expect?

#### Failure streaks follow a Weibull distribution

One key difference between progression and stagnation regimes is the propensity to reuse past components. From the perspective of exploration versus exploitation<sup>27,28</sup>, however, reuse helps one to retain a good version when it appears, but it could also keep one in a suboptimal position for longer, leading to our final prediction: the length of failure streaks follows a Weibull distribution (Supplementary Table 4):

$$P(N \ge n) \sim e^{-(n/\lambda)^{\beta}} \tag{2}$$

Moreover, the shape parameter  $\beta$  is connected with the temporal scaling exponent y through a scaling identity (Supplementary Information 3.8)

$$\beta + \gamma = 1 \tag{3}$$

This means that if we fit the streak length distribution in Fig. 1h-j to obtain the shape parameter  $\beta$ , it should relate to the temporal scaling exponent y, which is obtained from Fig. 3d-f. Comparing  $\beta$  and y measured independently across all three datasets shows consistency between our data and the scaling identity Eq. (3) (Supplementary Table 4).

We test the robustness of our results along several dimensions, arriving at broadly consistent conclusions (Methods, Extended Data Figs. 5-9, Supplementary Information 7). We include further quantitative tests for model assumptions and additional interpretations of the model in the Methods.

#### **Discussion**

As a single parameter, k necessarily combines individual, organizational and environmental factors in learning<sup>19,22</sup> (Supplementary Information 3.1). The one-parameter model developed here represents a minimal model (Supplementary Information 3.7), which can be extended into more complex frameworks. For example, agents may have varied incentives to improve or may differ in their confidence and ability to judge their previous work. Such factors trace heterogeneity in the population and can be captured by the  $\alpha$  parameter, which quantifies the propensity of individuals to change given feedback. This led us to develop the  $k-\alpha$  model (Methods), which predicts a two-dimensional phase diagram with three distinct phases (Extended Data Fig. 10a, b, Methods, Supplementary Information 4.1). The model can be further extended to capture fuzzy inference from past feedback, allowing agents to not always choose the best previous versions (see ' $k-\alpha-\delta$ 

model' in the Methods, Extended Data Fig. 10c, d, Supplementary Information 4.2).

The model also offers relevant insights for the understanding of learning curves. For example, the second critical point of the model suggests the existence of a minimum number of failures one needs to consider  $(k^*+1)$ , indicating that it is unnecessary to learn from all past experiences to achieve a maximal learning rate. This finding poses a potential explanation for the widespread nature of Wright's law across a wide variety of domains, particularly given the fact that in many of those domains not all past experiences can be considered (Supplementary Information 2).

Furthermore, our simple model does not explicitly account for many of the complexities that characterize real settings that may affect failure dynamics, such as knowledge depreciation<sup>29</sup>, competition, forgetting and transfer<sup>12</sup> or vicarious learning from others<sup>30</sup>. However, the model offers a theoretical basis to incorporate additional factors, including individual and organizational characteristics that may affect learning 12,17 (see Methods for various factors related to learning rate, including organizationallearning, previous achievements and gender differences), demonstrating that our modelling framework can serve as a springboard for anchoring future models and analyses.

#### **Concluding remarks**

Together, these results support the hypothesis that if future attempts systematically build on past failures, the dynamics of repeated failures may reveal statistical signatures that are discernible at an early stage. Traditionally the main distinction between ultimate success and failure following repeated attempts has been attributed to differences in luck, learning strategies or individual characteristics, but here our model offers an important explanation with crucial implications: Even in the absence of distinguishing initial characteristics, agents may still experience fundamentally different outcomes. Indeed, Thomas Edison once said, 'Many of life's failures are people who did not realize how close they were to success when they gave up.' Our results unveil identifiable early signals that help us to predict the eventual outcome to which failures lead. Together, they not only deepen our understanding of the complex dynamics beneath failure, but also hold lessons for individuals and organizations that experience failure and the institutions that aim to facilitate or hinder their eventual breakthrough.

#### **Online content**

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1725-y.

- Fortunato, S. et al. Science of science. Science 359, eaao0185 (2018)
- 2. Harford, T. Adapt: Why Success Always Starts with Failure (Farrar, Straus and Giroux, 2011).
- 3. Wuchty, S., Jones, B. F. & Uzzi, B. The increasing dominance of teams in production of knowledge. Science 316, 1036-1039 (2007).
- Jones, B. F. The burden of knowledge and the "death of the renaissance man": is innovation getting harder? Rev. Econ. Stud. 76, 283-317 (2009).
- Sinatra, R., Wang, D., Deville, P., Song, C. & Barabási, A.-L. Quantifying the evolution of individual scientific impact. Science 354, aaf5239 (2016).
- Liu. L. et al. Hot streaks in artistic, cultural, and scientific careers, Nature 559, 396-399

Hu, Y., Havlin, S. & Makse, H. A. Conditions for viral influence spreading through multiplex

- correlated social networks. Phys. Rev. X 4, 021031 (2014). Barabási, A.-L. The origin of bursts and heavy tails in human dynamics. Nature 435,
- 207-211 (2005)
- González, M. C., Hidalgo, C. A. & Barabási, A.-L. Understanding individual human mobility patterns, Nature 453, 779-782 (2008).
- Castellano, C., Fortunato, S. & Loreto, V. Statistical physics of social dynamics, Rev. Mod. Phys. 81, 591-646 (2009). Malmgren, R. D., Stouffer, D. B., Campanharo, A. S. & Amaral, L. A. N. On universality in
- human correspondence activity, Science 325, 1696-1700 (2009) Argote, L. Organizational Learning: Creating, Retaining and Transferring Knowledge
- (Springer Science & Business Media, 2012). Sitkin, S. B. Learning through failure: the strategy of small losses. Res. Organ. Behav. 14,
- 231-266 (1992) Yelle, L. E. The learning curve: historical review and comprehensive survey. Decis. Sci. 10,
- 302-328 (1979). Dutton, J. M. & Thomas, A. Treating progress functions as a managerial opportunity.
- Acad. Manage. Rev. **9**, 235-247 (1984). Huber, G. P. Organizational learning: the contributing processes and the literatures.
- Organ. Sci. 2, 88-115 (1991). Cannon, M. D. & Edmondson, A. C. Failing to learn and learning to fail (intelligently): how
- great organizations put failure to work to innovate and improve. Long Range Plann. 38, Kaplan, S. N. & Lerner, J. in Measuring Entrepreneurial Businesses: Current Knowledge
- and Challenges (Univ. Chicago Press, 2016). Eggers, J. P. & Song, L. Dealing with failure: serial entrepreneurs and the costs of
- changing industries between ventures, Acad. Manage, J. 58, 1785-1803 (2015).
- National Consortium for the Study of Terrorism and Responses to Terrorism. Global Terrorism Database (GTD) https://www.start.umd.edu/research-projects/global-terrorismdatabase-atd (2018).
- Clauset, A. & Gleditsch, K. S. The developmental dynamics of terrorist organizations. PLoS ONE 7, e48633 (2012).
- Johnson, N. et al. Pattern in escalations in insurgent and terrorist activity. Science 333. 22 81-84 (2011).
- Newell, A. & Rosenbloom, P. S. in Cognitive Skills and their Acquisition 1 (ed. Anderson, J. R.) 1-55 (Erlbaum, 1981).
- 24. Anderson, J. R. Acquisition of cognitive skill, Psychol, Rev. 89, 369-406 (1982)
- 25. Muth, J. F. Search theory and the manufacturing progress function. Manage. Sci. 32, 948-962 (1986).
- Wright, T. P. Factors affecting the cost of airplanes. J. Aeronaut. Sci. 3, 122-128 (1936).
- March, J. G. Exploration and exploitation in organizational learning. Organ. Sci. 2, 71-87 27. (1991).
- 28. Foster, J. G., Rzhetsky, A. & Evans, J. A. Tradition and innovation in scientists' research strategies. Am. Sociol. Rev. 80, 875-908 (2015).
- Arbesman, S. The Half-life of Facts: Why Everything We Know Has an Expiration Date
- Madsen, P. M. & Desai, V. Failing to learn? The effects of failure and success on organizational learning in the global orbital launch vehicle industry. Acad. Manage. J. 53, 451-476 (2010).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations

© The Author(s), under exclusive licence to Springer Nature Limited 2019

#### Methods

#### **Model assumptions**

Parameter k in our model can be viewed as approximating the 'memory' of past versions. The rationale of using k for the model is rooted in the learning literature showing that the general notion of 'forgetting' takes multiple forms, often representing a combination of individual, organizational and environmental factors. Indeed, several relevant factors may be at play, which can generate patterns similar to forgetting. For example, in rapidly shifting innovation domains, not all past failures remain useful over time and some become obsolete. Consider the possibility of knowledge depreciation<sup>31</sup>, which could also apply in our settings as environments (of scientific knowledge, capital markets or security situations) evolve over time, such that past experience could become useless even if memorized. For example, an NIH proposal four failures ago may become irrelevant as the ideas proposed have been proven wrong, or published by the principal investigator or another research group <sup>32,33</sup>. Similarly, startup ideas from the dot-com era may be irrelevant in the era of artificial intelligence and Blockchain<sup>34</sup>. Terrorist tactics can also depreciate over time, as past strategies attracted media coverage and gave rise to tighter security measures to defend against them<sup>22</sup>. This line of reasoning supports the intuition that recent attempts are most relevant. It is also consistent with the learning literature, which suggests knowledge forgetting can happen in distinct ways, either voluntarily or involuntarily<sup>35</sup>. Given these factors, here we select a single parameter k to encapsulate a variety of potential contributing factors.

#### Quantifying component dynamics

To empirically measure the dynamics of components, we collected abstract information for all R01 applications submitted after 2008 (Supplementary Information 5.4). To this data corpus we applied a natural-language-processing technique to extract MeSH (medical subject headings) terms from each abstract, which approximate the methods, physical states and processes involved in the proposed research. This allows us to quantify the dynamics of component reuse from previous proposals for the successful group. We measure the new versions of components by the number of new MeSH terms (terms that did not appear in the previous k submissions, defined as  $m_n$ ) and plot  $M_n \equiv \langle m_n \rangle / \langle m_1 \rangle$  as a function of n. Our model suggests that given k, we can use  $M_n$  to mimic the temporal dynamics of  $T_n$ . More precisely, for the successful group, we expect to observe that for large  $k(k > k^*)$ ,  $M_n$  and  $T_n$  are characterized by similar dynamics. For small k ( $k < k^*$ ), however, the two quantities could be quite different. As shown in Extended Data Fig. 4, our empirical analysis shows that the two curves indeed follow different dynamics for small k ( $k \le 3$ ), but the dynamics of  $M_n$  and  $T_n$  become statistically indistinguishable for k>3 (from 4 to ∞), approximately following a power law with y ~ 0.35. We cannot directly examine component dynamics for the unsuccessful group due to the lack of sufficient data—by definition agents in this group submitted no proposal after 2010, and the unsuccessful abstract data only go back to 2008.

#### **Phase transitions**

To understand the nature of two transition points in our model, here we consider a canonical ensemble of N particles  $(N \to \infty)$  and three energy states  $E_a(h) = 1$ ,  $E_b(h) = (2h-1)^2$  and  $E_c(h) = 1$  where h denotes the external field. We can write down the partition function of the system  $Z = e^{-NE_a(h)} + e^{-NE_b(h)} + e^{-NE_c(h)}$ , and calculate its free energy density  $f = \ln[Z/N]$ . In this system, it can be shown that the magnetization density  $m = \frac{df}{dh}$  is discontinuous at the boundary of two energy states  $E_a(h) = E_b(h)$  and  $E_b(h) = E_c(h)$ , characterized by two phase transitions at h = 0 and h = 1, respectively.

We notice that the canonical ensemble considered above has a mapping to our model. Indeed, denoting  $\Gamma = k^* \times \gamma/(1-\gamma)$  and  $K = k - k^*$ , we can rescale the system as  $\Gamma = \min\{\max\{\Gamma_a(K), \Gamma_b(K)\}, \Gamma_c(K)\}$  where

 $\Gamma_a(K) = 0$ ,  $\Gamma_b(K) = K$  and  $\Gamma_c(K) = 1$ , allowing us to map the two systems through  $f \to (2\Gamma - 1)^2$ ,  $N \to \ln[n]$ ,  $h \to K$  and  $E_i(h) = (2\Gamma_i(K) - 1)^2$  (Extended Data Fig. 1g–j).

To understand the origin of the two transition points, we can calculate the expected lifespan of a high-quality version, obtaining  $\langle u(x)\rangle \sim \langle (1-x)^{-\min[k/k^*,1/k^*+1]}\rangle$  (Supplementary Information 3.4). The first critical point  $k^*$  occurs when the first moment  $\langle u\rangle$  diverges. Indeed, when k is small  $(k < k^*)$ ,  $\langle u\rangle$  is finite, indicating that high-quality versions can only be reused for a limited period of time. Once k passes critical point  $k^*$ , however,  $\langle u\rangle$  diverges, offering the possibility for a high-quality version to be retained for an unlimited period of time. The second critical point arises due to the competition between two dynamical forces: (1) whether the current best version becomes forgotten after k consecutive attempts in creating new versions (dominated by the  $k/k^*$  term); or (2) it is substituted by an even better version (dominated by the  $1/k^*+1$  term).

Note that while phase transitions carry exceptional importance in statistical physics, similar phenomena and concepts are also of fundamental relevance in the social and behavioural science literatures. For example, critical thresholds have been observed and modelled in social settings that include shifts in the segregation of neighbourhoods<sup>36</sup>, the formation of social networks<sup>37</sup> and changes in collective opinions<sup>38</sup>. In each case, slight shifts in microscale phenomena, such as average preference, group size or interaction intensity, condition a qualitative transition in macroscale outcomes.

#### Alternative hypothesis, interpretation and robustness checks

To better understand the role of heterogeneity in learning, we separated the successful group into narrow-win and clear-win subgroups based on their eventual performance. We find that, despite their eventual difference, the temporal dynamics of the two groups remain statistically indistinguishable (two-sided Welch's t-test, P = 0.763 ( $D_1$ ), 0.813 ( $D_2$ ), 0.259 ( $D_3$ ), Extended Data Fig. 4), suggesting that the distinction between successful and unsuccessful groups appears the most critical, whereas agents within the successful group are characterized by similar dynamics, consistent with the predictions of our model.

An alternative interpretation for the stalled efficiency of the unsuccessful group is an effort to hedge against failures—their efficiency did not improve because they spent more effort elsewhere. The three professions that we studied, NIH investigators, entrepreneurs and terrorists, involve varied levels of risk, exposure and commitment, which renders this explanation less likely.

To test the robustness of our results, we vary the definitions of what constitutes the successful group (Supplementary Information 7.1) by excluding revisions in  $D_1$  (Extended Data Fig. 6), changing the threshold of high-value mergers and acquisitions or controlling for unicorn companies in  $D_2$  (Extended Data Fig. 7), and varying the types of attack or changing the threshold for fatal attacks in  $D_3$  (Extended Data Fig. 8). We also vary the definition of unsuccessful groups (Extended Data Fig. 5, Supplementary Information 7.2) and test other measures to approximate performance (Extended Data Fig. 9j–o, Supplementary Information 7.4, 7.5). We further adjust for temporal variation by controlling for the overall success rate across different years (Extended Data Fig. 9a–i, Supplementary Information 7.3). Across all variations, our conclusions remain the same.

#### **Predicting ultimate success**

We use a simple logistic model to predict whether one may achieve success following N previously failed attempts in  $D_1$ , using only temporal features  $t_n$  ( $1 \le n \le N-1$ ) as predictors. To evaluate prediction accuracy, we calculate the area under the receiver operating characteristic (AUC) curve with tenfold cross-validation. We find that, by observing the timing of the first three failures alone, our simple temporal feature yields high accuracy in predicting the eventual outcome with an AUC close to 0.7, which is significantly higher than random guessing (Mann–Whitney U-test,  $P < 10^{-180}$ ; Extended Data Fig. 3a, Supplementary

Information 6.1). We repeated the same prediction task on  $D_2$  and  $D_3$ , arriving at similar conclusions (Extended Data Fig. 3b, c, Supplementary Information 6.1). The predictive power from temporal features alone is somewhat unexpected. Indeed, there are a large number of documented factors that affect the outcome of a grant application  $^{39-43}$ , ranging from the previous success rate to publication and citation records to the race, ethnicity and gender of the applicant. Here we ignore these factors, however, using only features that pertain to temporal scaling as prescribed by our model. This suggests that our predictive power represents a lower bound, which could be further improved and leveraged by incorporating additional factors.

#### k-α model

Agents may differ in the judgment of their own work or incentives to change given feedback, which can be captured by varying the a parameter in the original k model. Of the many influences on p, one key factor is the quality of existing versions, suggesting that p should be a function of  $x^*$ . Consider the following two extreme cases. If  $x^* o 0$ , existing versions of this component have one of the worst scores and, hence, a high potential for improvement when replaced with a new version. In this case, the likelihood of creating a new version is high, that is,  $p \rightarrow 1$ . On the other hand,  $x^* \rightarrow 1$  corresponds to a near-perfect version, yielding a decreased incentive to create a new one  $(p \to 0)$ . Indeed,  $P(x \ge x^*) = 1 - x^*$ captures the potential to improve on previous versions, prompting us to assume that  $p = (1 - x^*)^{\alpha}$  where  $\alpha > 0$  characterizes the propensity of an agent to create new versions given the quality of existing ones. Therefore,  $\alpha \rightarrow 0$  indicates that regardless of one's evaluation, the agent will always create a new version, whereas  $\alpha \rightarrow \infty$  points to the other extreme where one does not create a new version unless it is extremely bad (Extended Data Fig. 10a). Considering  $\alpha$  another tunable parameter, we arrive at a two-parameter model: the  $k-\alpha$  model (Supplementary Information 4.1).

To solve this model we can substitute  $k^*$  with  $1/\alpha$ , and the indexes  $k/k^*$  and  $1/k^*+1$  now become  $k\alpha$  and  $\alpha+1$ . The extended model thus predicts the existence of three different phases on a two-dimensional phase diagram, with boundaries  $k\alpha=1$  and  $(k-1)\alpha=1$  that separate the three phases (Extended Data Fig. 10b). The  $k-\alpha$  model reduces back to the two critical points in the original k model when we fix  $\alpha$ . The two parameters jointly define an effective  $K \equiv k-k^*=k-1/\alpha$ . The critical boundaries therefore reduce into two simple equations: K=0 and K=1. Note that the assumed relationship between p and  $(1-x^*)$  is not limited to a power law but can be relaxed into its asymptotic form. Indeed, we show that as long as the function satisfies  $\frac{\ln |p|}{\ln (1-x^*)} \rightarrow \alpha$  as  $x^* \rightarrow 1$ , the model offers the same predictions  $x^{25}$  (Extended Data Fig. 3, Supplementary Information 3.6).

#### $k-\alpha-\delta$ model

Agents may have fuzzy or unclear inference regarding past feedback, and may therefore not always choose the version with highest quality. We can model the choice between different versions in a probabilistic fashion, by introducing a  $\delta$  parameter to the  $k-\alpha$  model. Here the probability to choose the ith version as a baseline follows

$$P(i) = \frac{1}{Z} (1 - x_i)^{-\delta} 1_{n - k \le i \le n - 1}$$

where Z is the normalization factor,  $Z = \sum_{i=n-k}^{n-1} (1-x_i)^{-\delta}$  and  $k \ge 1$ .  $\delta = 0$  means one cannot differentiate between the quality of past versions and selects randomly among different versions, whereas  $\delta \to \infty$  indicates that one always chooses the previous version with highest quality, converging back to our original k model or the  $k-\alpha$  model. Incorporating  $\delta$  leads to the  $k-\alpha-\delta$  model (Supplementary Information 4.2).

Analytically solving the model reveals interesting scaling behaviours based on  $\delta$  (Supplementary Information 4.2). Indeed, we find the scaling behaviour of the system follows

$$y(k, \alpha, \delta) = 1 - \{\max[\min(\alpha + (k-1)\min\{1, \alpha, \delta\}, \alpha + 1), 1]\}^{-1}$$

with rich mathematical properties. When  $\delta \to \infty$ , the new solutions converge back to the original solution for the  $k-\alpha$  model. With  $\delta$ , the three-parameter model is characterized by four different phases. Three of the regimes are generalizations of those found in the  $k-\alpha$  model, where the scaling exponent  $\gamma$  does not depend on  $\delta$  in the limit of  $\delta \to \infty$ , that is,  $\gamma(k,\alpha,\delta)=\gamma(k,\alpha,\infty)$ . The fourth one, however, is a new phase and only exists for small  $\delta$ . The intuition is that in this regime the inability to select a high-quality version (small  $\delta$ ) dominates the scaling behaviour, with exponent  $\gamma(k,\alpha,\delta)=1-[(k-1)\delta+\alpha]^{-1}$ . Together, these extensions offer further support for the predictions of our original model, while demonstrating the theoretical potential of the model by enriching its mathematical properties for more realistic interpretations. They also point to promising future research that explores the interplay between different perspectives on learning.

Note that although all three variations of the model predict the existence of different phases, the primary focus of this paper concerns the fundamental differences in the nature of these regimes (that is, stagnation versus progression), rather than the behaviour of the system as it approaches the critical threshold. As such, the conclusions of the paper hold the same regardless of any specific critical behaviour around the threshold.

#### Factors related to learning rate

Our model offers a framework to anchor potential factors relevant to learning<sup>44,45</sup>. As an example, here we test three different factors. First, the literature has identified several factors for the emergence of learning at the level of organizations  $^{12,21}$ , suggesting that individual learning is just one factor in how and why organizations learn. This suggests that settings closer to organizational learning (such as terrorist groups) should correspondingly experience higher learning rates than those closer to individual learning (such as NIH principal investigators) (Supplementary Information 5.5). We test this hypothesis by calculating the average scaling exponent y measured from our data (Supplementary Table 4), and find that our estimations support this hypothesis; learning rates are lowest for individual researchers, higher for entrepreneurs and their founding teams and highest for terrorist organizations. Note that although these results show consistency with theories from the organizational learning literature, these differences could also be due to inherent domain-specific differences.

Second, higher previous achievements often bring recognition and resources, a phenomena referred to as the Matthew effect <sup>46</sup>, which might translate into higher learning rates. To test this we link NIH grant application data to the Web of Science citation database through a systematic effort to disambiguate authors, and match the citations of previous research papers with submitted proposals <sup>5,47</sup> (Supplementary Information 5.6). We take principal investigators who failed more than three times before their eventual success and calculate the total number of citations from all his/her papers including only papers published before the first failure. We find that prior acclaim is positively and significantly correlated with learning rate  $\gamma$  (P<0.001).

Third, persistent gender inequalities in science and entrepreneur-ship  $^{48-50}$  suggest the possibility that failure dynamics may be mediated by gender  $^{51,52}$ . Our regression analysis reveals a significant correlation between gender and learning rate (Supplementary Information 5.7). All else being equal, the learning rate  $\gamma$  of a male principal investigator in the NIH system exceeds that of a female principal investigator by 0.14 ( $P\!=\!0.001$ ), suggesting that male principal investigators fail faster than their female colleagues. This difference appears substantial, considering that the average learning rate is centred around 0.35. We further test this relationship in the startup dataset, finding a similar gap of 0.10 between male and female innovators, but this result is not significant, possibly owing to a smaller sample size. Note that these gender differences probably flow from institutional as well as individual causes, such

as a culture that discourages women from persistence and encourages oversensitivity to feedback. Indeed, one irony suggested by our model is that agents in the stagnation region did not work less. Rather they made more, albeit unnecessary modifications to what were otherwise advantageous experiences.

#### **Reporting summary**

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

#### **Data availability**

This paper makes use of restricted access data from the National Institutes of Health (NIH), protected by the Privacy Act of 1974 as amended (5 U.S.C. 552a). Deidentified data necessary to reproduce all plots and statistical analyses are freely available at https://yian-yin.github.io/quantifyFailure. Those wishing to access the raw data can apply for access following the procedures outlined in the NIH Data Access Policydocument(http://report.nih.gov/pdf/DataAccessPolicy.pdf). The VentureXpert database is available from Thomson Reuters. The Global Terrorism Database is publicly available at https://www.start.umd.edu/gtd/.

#### **Code availability**

Code is available at https://yian-yin.github.io/quantifyFailure.

- Argote, L., Beckman, S. L. & Epple, D. The persistence and transfer of learning in industrial settings. Manage. Sci. 36. 140–154 (1990).
- 32. Kuhn, T. S. The Structure of Scientific Revolutions (Chicago Univ. Press, 2012).
- Merton, R. K. Singletons and multiples in scientific discovery: a chapter in the sociology of science. Proc. Am. Phil. Soc. 105, 470–486 (1961).
- Gompers, P., Kovner, A., Lerner, J. & Scharfstein, D. Performance persistence in entrepreneurship. J. Financ. Econ. 96, 18–32 (2010).
- de Holan, P. M. & Phillips, N. Remembrance of things past? the dynamics of organizational forgetting. Manage. Sci. 50, 1603–1613 (2004).
- 36. Schelling, T. C. Micromotives and Macrobehavior (WW Norton & Company, 2006).
- Watts, D. J. A simple model of global cascades on random networks. Proc. Natl Acad. Sci. USA 99, 5766–5771 (2002).
- Holme, P. & Newman, M. E. Nonequilibrium phase transition in the coevolution of networks and opinions. *Phys. Rev. E* 74, 056108 (2006).
- Ginther, D. K. et al. Race, ethnicity, and NIH research awards. Science 333, 1015–1019 (2011).

- Boudreau, K. J., Guinan, E. C., Lakhani, K. R. & Riedl, C. Looking across and looking beyond the knowledge frontier: intellectual distance, novelty, and resource allocation in science. Manage. Sci. 62, 2765–2783 (2016).
- Bromham, L., Dinnage, R. & Hua, X. Interdisciplinary research has consistently lower funding success. *Nature* 534, 684–687 (2016).
- Banal-Estanol, A., Macho-Stadler, I. & Pérez Castrillo, D. Key Success Drivers in Public Research Grants: Funding the Seeds of Radical Innovation in Academia? CESifo Working Paper Series 5852 (CESifo, 2016).
- Ma, A., Mondragón, R. J. & Latora, V. Anatomy of funded research in science. Proc. Natl Acad. Sci. USA 112, 14760–14765 (2015).
- 44. Levitt, B. & March, J. G. Organizational learning. Annu. Rev. Sociol. 14, 319-338 (1988).
- 45. Argote, L. & Epple, D. Learning curves in manufacturing. Science 247, 920-924 (1990).
- 46. Merton, R. K. et al. The Matthew effect in science. Science 159, 56-63 (1968).
- Huang, J., Ertekin, S. & Giles, C. L. Efficient name disambiguation for large-scale databases. In European Conference on Principles of Data Mining and Knowledge Discovery 536–544 (Springer, 2006).
- 48. Shen, H. Inequality quantified: Mind the gender gap. Nature 495, 22-24 (2013).
- Larivière, V., Ni, C., Gingras, Y., Cronin, B. & Sugimoto, C. R. Bibliometrics: global gender disparities in science. *Nature* 504. 211–213 (2013).
- Yang, T. & Aldrich, H. E. Who's the boss? Explaining gender inequality in entrepreneurial teams. Am. Sociol. Rev. 79, 303–327 (2014).
- Argote, L., Insko, C. A., Yovetich, N. & Romero, A. A. Group learning curves: the effects of turnover and task complexity on group performance. J. Appl. Soc. Psychol. 25, 512– 529 (1995).
- Bailey, C. D. Forgetting and the learning curve: a laboratory study. Manage. Sci. 35, 340– 352 (1989).

**Acknowledgements** We thank C. Song, A. Clauset, B. Uzzi, B. Jones, E. Finkel, J. Van Mieghem, A. Bassamboo and Y. Xie for helpful discussions, and H. Sauermann and S. Havlin for suggesting extensions of the model, leading us to discover the k- $\alpha$  and k- $\alpha$ - $\delta$  models. This work is supported by the Air Force Office of Scientific Research under award number FA9550-15-1-0162, FA9550-17-1-0089 and FA9550-19-1-0354, National Science Foundation grant SBE 1829344, the Alfred P. Sloan Foundation G-2019-12485, and Northwestern University Data Science Initiative. This work does not reflect the position of NIH.

**Author contributions** D.W. conceived the project and designed the experiments; Y.Y. and Y.W. collected data and performed empirical analyses with help from D.W. and J.A.E.; Y.Y. and D.W. carried out theoretical calculations; all authors collaboratively designed the model and interpreted results; D.W. and Y.Y. wrote the manuscript; all authors edited the manuscript.

**Competing interests** Y.W. and D.W. serve as special volunteers (unpaid) to the NIH. The remaining authors declare no competing interests.

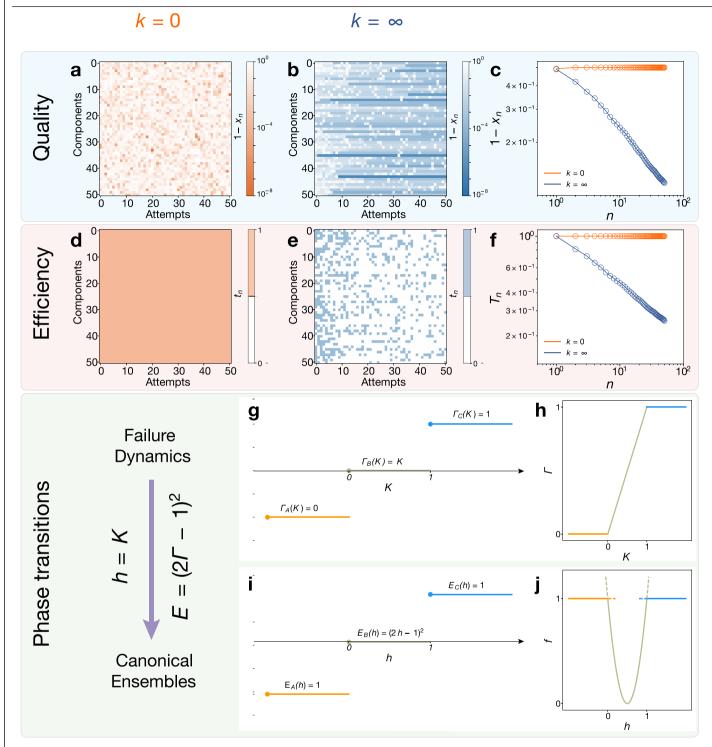
#### Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-

 $\textbf{Correspondence and requests for materials} \ \text{should be addressed to D.W.}$ 

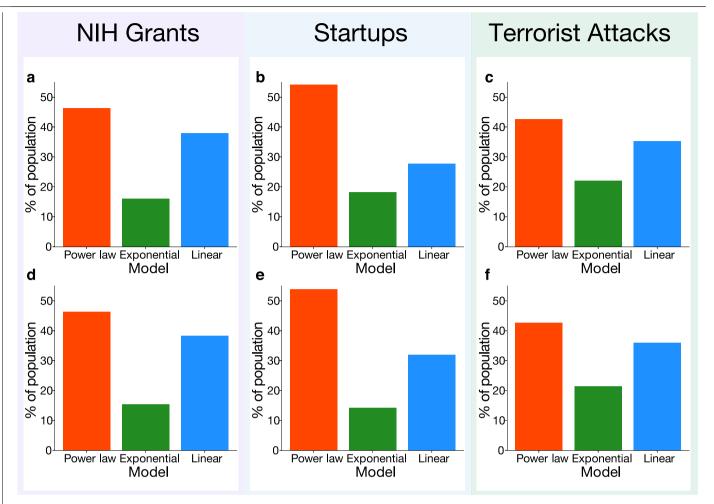
 $\label{per review information Nature} \ \ thanks \ Shlomo \ Havlin \ and \ Henry \ Sauermann \ for \ their contribution to the peer review of this work.$ 

Reprints and permissions information is available at http://www.nature.com/reprints.



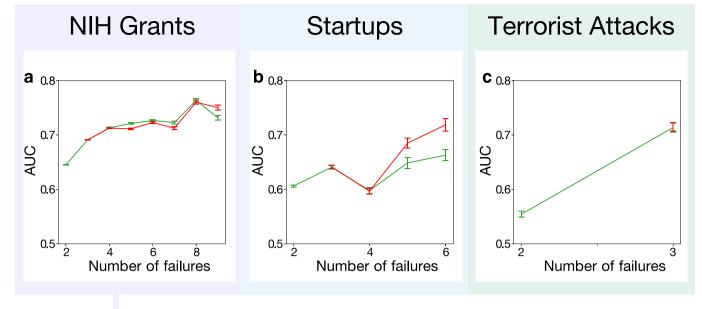
**Extended Data Fig. 1**| **The** k **model. a-f**, Simulation results from the model  $(\alpha = 0.6)$  for the cases of k = 0 (**a**, **d**) and  $k \to \infty$  (**b**, **e**) in terms of the average quality (**a-c**) and efficiency (**d-f**) of each attempt. k = 0 recovers the chance model, predicting a constant quality (**c**) and efficiency (**f**).  $k \to \infty$  predicts temporal scaling that characterizes the dynamics of failure (**e**) with improved quality (**b**), recovering predictions from learning curves and Wright's law. **g-j**, Illustration of mapping between failure dynamics (**g**, **h**) and canonical ensembles (**i**, **j**). The

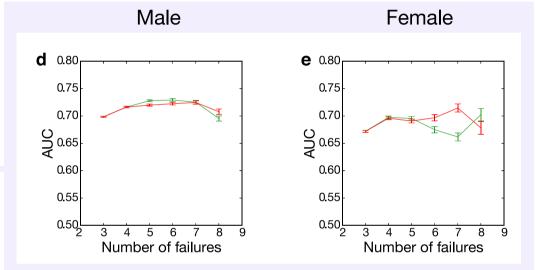
canonical system is characterized by three different states a, b, c with corresponding energy densities  $E_a(h)$ ,  $E_b(h)$ ,  $E_c(h)$ . Here we assume  $E_a(h) = (2\varepsilon h - 1)^2$ ,  $E_b(h) = (2h - 1)^2$  and  $E_c(h) = [2\varepsilon (1 - h) - 1]^2$  where  $\varepsilon \to 0^*$ . The introduction of  $\varepsilon$  is to distinguish state a from state c, both of which can be approximated in the limiting condition  $E_a(h) = E_c(h) = 0$ . We map  $f \to (2\Gamma - 1)^2$ ,  $N \to \ln[n]$ ,  $h \to K$  and  $E_i(h) = [2\Gamma_i(K) - 1]^2$ . In this case, the two transition points  $k^*$  and  $k^* + 1$  correspond to h = 0 and 1 in the canonical ensemble systems.



**Extended Data Fig. 2** | **Predicting temporal dynamics in science, entrepreneurship and security.**  $\mathbf{a}$ - $\mathbf{c}$ , We compare the goodness of fit for three different models in temporal dynamics in NIH grants ( $\mathbf{a}$ , n = 10345), startups ( $\mathbf{b}$ , n = 275) and terrorist attacks ( $\mathbf{c}$ , n = 136). For each individual sample, we take all but the last inter-event time for model fitting (n = 1, ..., N - 1), comparing model

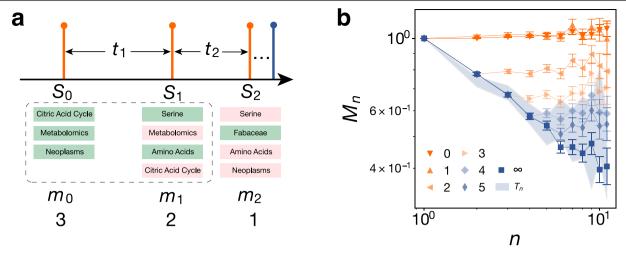
predictions for the last inter-event time. The tested functional forms are power law,  $t_n = an^b$ ; exponential,  $t_n = ab^{-n}$ ; and linear,  $t_n = a + bn$ . We then calculate the frequency that each model reaches minimum error, defined as  $|\log(t_N) - \log(\hat{t}_N)|$ , among all three forms. The power-law model offers consistently better predictions.  $\mathbf{d} - \mathbf{f}$ , As in  $\mathbf{a} - \mathbf{c}$ , but using  $|t_N - \hat{t}_N|$  as the loss function.

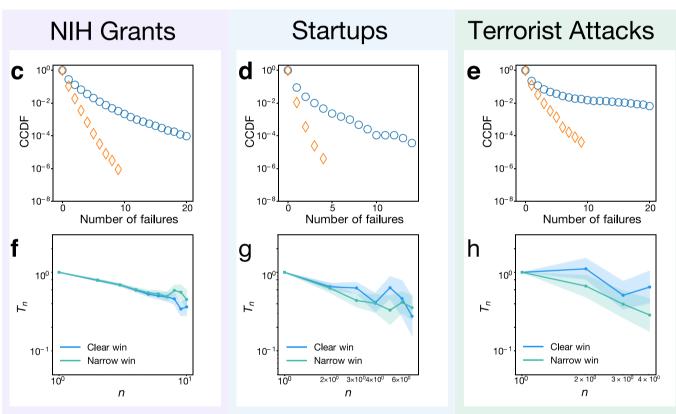




Extended Data Fig. 3 | Predicting ultimate success in science, entrepreneurship and security. a-c, Area under the receiver operating characteristic curve (AUC) of the prediction task. We apply two logistic regression models (Supplementary Information 6.1) to predict ultimate success in NIH grants (a), startups (b) and terrorist attacks (c). The centres and error bars

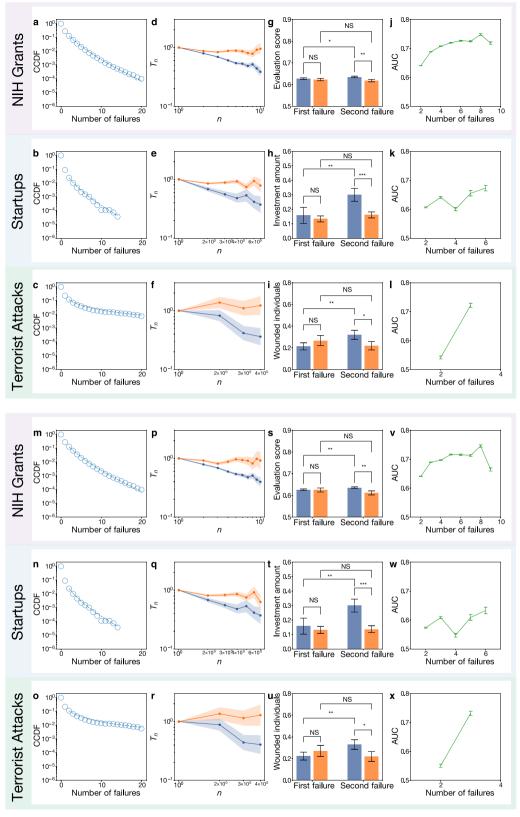
of AUC scores denote the mean  $\pm$  s.e.m. calculated from tenfold cross-validation over 50 randomized iterations (green, model 1; red, model 2). **d, e**, As in **a** but predicting ultimate success in NIH grants for male (**d**) and female (**e**) investigators.





**Extended Data Fig. 4** | **Model validations. a**, **b**, An illustration of the component dynamics. We extract all MeSH terms associated with the nth attempt,  $S_n$ , and calculate the number of new terms  $m_n$ , defined as  $|S_n - (S_{n-1} \cup \cdots \cup S_{n-k})|$ . **b**, Testing component dynamics in NIH grant applications. We calculate the dynamics of  $M_n = \langle m_n \rangle / \langle m_1 \rangle$  using different k and compare it with  $T_n$ . The centres and error bars of  $M_n$  show the mean  $\pm$  s.e.m. (n = 5,899) for different k. The shaded area shows mean  $\pm$  s.e.m. of  $T_n$  (log scale) measured on the same subset. All k > 3 lead to similar trends between  $M_n$  and  $T_n$ .  $\mathbf{c} - \mathbf{e}$ , Length of failure streak after randomization in science ( $\mathbf{c}$ ),

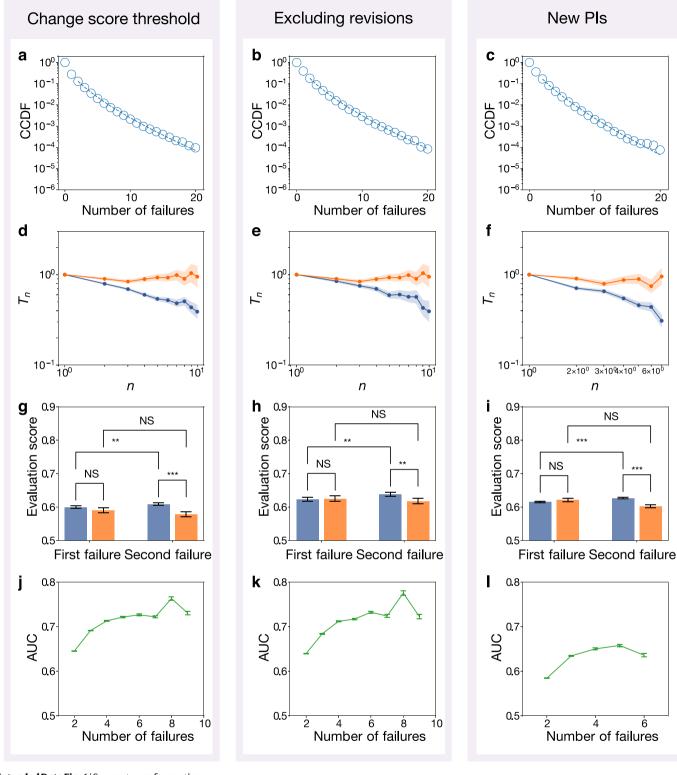
entrepreneurship (**d**) and security (**e**). We take the samples used in Fig. 1 and shuffle the success/failure label from each attempt. This operation keeps both the overall success rate and the total number of attempts for each individual constant. **f**-**h**, Temporal scaling patterns within the successful group in science (**f**), entrepreneurship (**g**) and security (**h**). We separated the successful group into two subgroups (narrow winners and clear winners) based on eventual performance (0.9 in evaluation score for  $D_1$ , 0.5 in investment amount for  $D_2$  and 1 in wounded individuals for  $D_3$ ). The shaded area shows mean  $\pm$  s.e.m. of  $T_n$  (log scale).



**Extended Data Fig. 5** | See next page for caption.

**Extended Data Fig. 5** | **Robustness check on definition of unsuccessful group. a–1**, Robustness check as we change the threshold of inactivity to 3 years. **a–c**, Failure streak inscience (**a**), entrepreneurship (**b**) and security (**c**). Blue circles represent real data from the successful group and dashed lines represent fitted Weibull distributions. **d–f**, Temporal scaling patterns in science (**d**), entrepreneurship (**e**) and security (**f**). The shaded area shows mean  $\pm$  s.e.m. of  $T_n$  (log scale). **g–i**, Performance dynamics in science (**g**, n = 641, 231, 578, 190, from left to right), entrepreneurship (**h**, n = 248, 1,332, 237, 1,312 from left to right) and security (**i**, n = 238, 198, 236, 199, from left to right). The successful and unsuccessful groups that experienced a large number of consecutive failures before the last attempt (at least 5 for  $D_1$ , 3 for  $D_2$  and 2 for  $D_3$ ) appear indistinguishable for first failures (two-sided Welch's t-test; P = 0.566, 0.671 and 0.349), but quickly diverge for second failures (two-sided Welch's t-test;

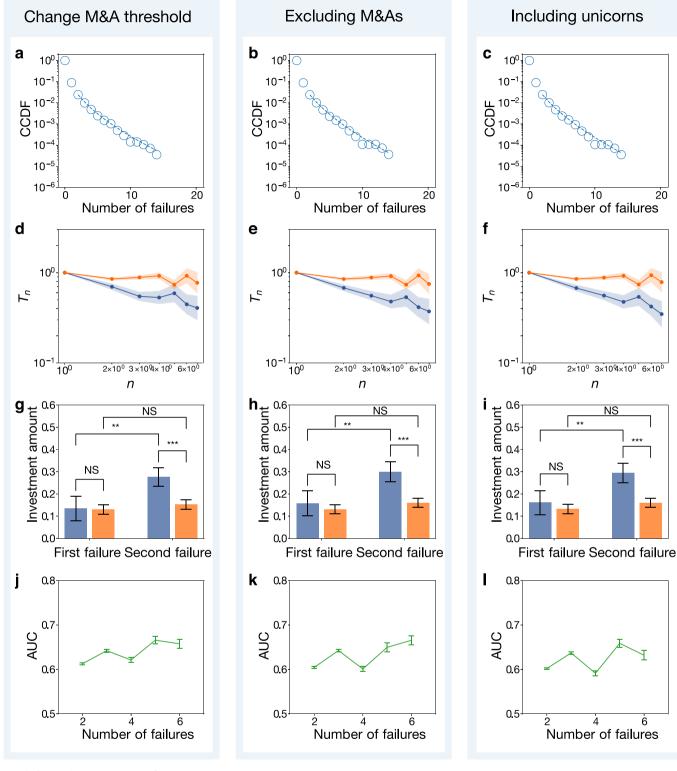
 $P=2.09\times10^{-2},4.95\times10^{-3}$  and  $7.77\times10^{-2}$ ). The successful group also shows significant improvement in performance (one-sided Welch's t-test;  $P=7.03\times10^{-2},2.37\times10^{-2}$  and  $2.32\times10^{-2}$ ), which is absent for the unsuccessful group (one-sided Welch's t-test; P=0.717,0.176 and 0.786). Data are mean  $\pm$  s.e.m.  $\mathbf{j}$ –I, AUC score of predicting ultimate success in science ( $\mathbf{j}$ ), entrepreneurship ( $\mathbf{k}$ ) and security ( $\mathbf{l}$ ). The centres and error bars of AUC scores denote the mean  $\pm$  s.e.m calculated from tenfold cross-validation over 50 randomized iterations.  $\mathbf{m}$ – $\mathbf{x}$ , As in  $\mathbf{a}$ –I but using 7 years as the threshold of inactivity. Sample sizes are  $\mathbf{s}$ : n=620,101,559,76;  $\mathbf{t}$ : n=248,977,237,989;  $\mathbf{u}$ : n=216,152,214,153. Pvalues in  $\mathbf{s}$ – $\mathbf{u}$  (from bottom to top) are P=0.883 ( $\mathbf{s}$ ), 0.671 ( $\mathbf{t}$ ), 0.456 ( $\mathbf{u}$ );  $P=2.25\times10^{-2}$  ( $\mathbf{s}$ ),  $1.38\times10^{-3}$  ( $\mathbf{t}$ ),  $8.34\times10^{-2}$  ( $\mathbf{u}$ );  $P=4.59\times10^{-2}$  ( $\mathbf{s}$ ),  $2.37\times10^{-2}$  ( $\mathbf{t}$ ),  $3.33\times10^{-2}$  ( $\mathbf{u}$ ); P=0.838 ( $\mathbf{s}$ ), 0.446 ( $\mathbf{t}$ ), 0.775 ( $\mathbf{u}$ ). \*P<0.1, \*\*P<0.05, \*\*\*P<0.01, NS, not significant ( $P\ge0.1$ ).



 $\textbf{Extended Data Fig. 6} \, | \, \textbf{See next page for caption.}$ 

**Extended Data Fig. 6 | Robustness check on**  $D_1$ **.** a**-c**, Failure streak as we change the score threshold to 55 (a), exclude revisions as successes (b) and only focus on new principal investigators without previous R01 grants (c). Blue circles represent real data from successful groups and dashed lines represent fitted Weibull distributions. d-f, Temporal scaling patterns as we change the score threshold to 55 (d), exclude revisions as successes (e) and only focus on new principal investigators without previous R01 grants (f). The shaded area shows mean  $\pm$  s.e.m. of  $T_n$  (log scale). g-i, Performance dynamics as we change the score threshold to 55 (g, n = 768, 189, 686, 170, from left to right), exclude revisions as successes (h, n = 252, 145, 216, 123, from left to right) and only focus on new principal investigators without previous R01 grants (i, n = 1,164, 308, 1,530, 334, from left to right). The successful and unsuccessful groups that experienced a

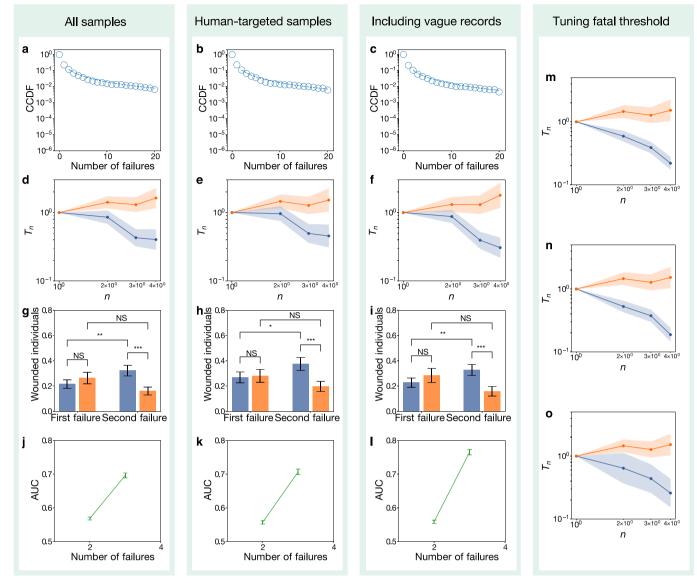
large number of consecutive failures before their last attempt (at least 5 for g and h, and 3 for i) appear indistinguishable for first failures (two-sided Welch's t-test; P = 0.242, 0.819, 0.289) but quickly diverge for second failures (two-sided Welch's t-test;  $P = 3.40 \times 10^{-4}$ ,  $3.40 \times 10^{-2}$ ,  $9.70 \times 10^{-7}$ ). The successful group also shows a significant improvement in performance (one-sided Welch's t-test;  $P = 4.23 \times 10^{-2}$ ,  $3.04 \times 10^{-2}$ ,  $1.92 \times 10^{-4}$ ), which is absent for the unsuccessful group (one-sided Welch's t-test; P = 0.863, 0.754, 0.997). Data are mean  $\pm$  s.e.m.  $\mathbf{j}$ –1, AUC score of predicting ultimate success as we change the score threshold to 55 ( $\mathbf{j}$ ), exclude revisions as successes ( $\mathbf{k}$ ) and only focus on new principal investigators without previous RO1 grants ( $\mathbf{l}$ ). The centres and error bars of AUC scores denote the mean  $\pm$  s.e.m calculated from tenfold cross-validation over 50 randomized iterations. \*P < 0.1, \*\*P < 0.05, \*\*\*P < 0.01, NS,  $P \ge 0.1$ .



 $\textbf{Extended Data Fig. 7} | See \ next \ page \ for \ caption.$ 

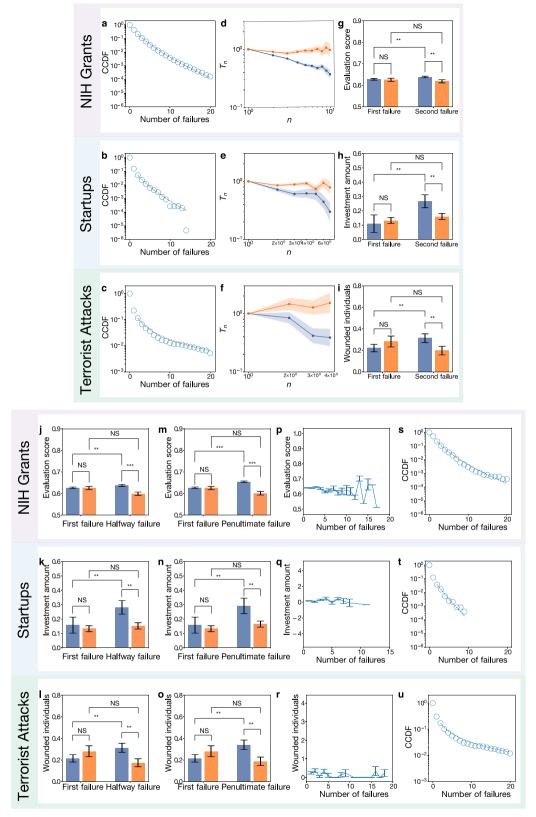
**Extended Data Fig. 7** | **Robustness check on**  $D_2$ . a-c, Failure streak as we change the threshold of high-value mergers and acquisitions (M&A) to 5% (a), exclude M&As as successes (b) and classify unicorns as successes (c). Blue circles represent real data from successful groups and dashed lines represent fitted Weibull distributions. d-f, Temporal scaling patterns as we change the threshold of high-value M&A to 5% (d), exclude M&As as successes (e) and include unicorns as successes (f). The shaded area shows mean  $\pm$  s.e.m. of  $T_n$  (log scale). g-f, Performance dynamics as we change the threshold of high-value M&A to 5% (g, n = 251, 1,304, 243, 1,284, from left to right), exclude M&As as successes (f), f = 248, 1,335, 237, 1,315, from left to right). The successful and unsuccessful groups that experienced a large number of consecutive failures before their last

attempt (at least 3) appear indistinguishable for first failures (two-sided Welch's t-test; P = 0.937, 0.647, 0.620) but quickly diverge for second failures (two-sided Welch's t-test; P = 9.92 ×  $10^{-3}$ , 4.94 ×  $10^{-3}$ , 6.33 ×  $10^{-3}$ ). The successful group also shows a significant improvement in performance (one-sided Welch's t-test; P = 2.16 ×  $10^{-2}$ , 2.37 ×  $10^{-2}$ , 2.77 ×  $10^{-2}$ ), which is absent for the unsuccessful group (one-sided Welch's t-test; P = 0.224, 0.158, 0.167). Data are mean  $\pm$  s.e.m. $\mathbf{j}$ - $\mathbf{l}$ , AUC score for predicting ultimate success as we change threshold of high-value M&A to 5% ( $\mathbf{j}$ ), exclude M&As as successes ( $\mathbf{k}$ ) and include unicorns as successes ( $\mathbf{l}$ ). The centres and error bars of AUC scores denote the mean  $\pm$  s.e.m calculated from tenfold cross-validation over 50 randomized iterations. \*P<0.1, \*\*P<0.05, \*\*\*P<0.01, NS, P<2 0.1.



**Extended Data Fig. 8** | **Robustness check on**  $D_3$ **. a-c**, Failure streak as we focus on all samples (a), samples of human-targeted attacks (b) and include vague data on fatalities (c). Blue circles represent real data from successful groups and dashed lines represent fitted Weibull distributions. **d-f**, Temporal scaling patterns as we focus on all samples (**d**), samples of human-targeted attacks (**e**) and include vague data on fatalities (**f**). The shaded area shows mean  $\pm$  s.e.m. of  $T_n$  (log scale). **g-i**, Performance dynamics as we focus on all samples (**g**, n = 231, 231, 229, 232, from left to right), samples of human-targeted attacks (**h**, n = 176, 173, 173, 174, from left to right) and include vague data on fatalities (**i**, n = 227, 147, 225, 148, from left to right). The successful and unsuccessful groups that experienced a large number of consecutive failures before their last attempt (at least 2) appear indistinguishable for first failures (two-sided Welch's t-test;

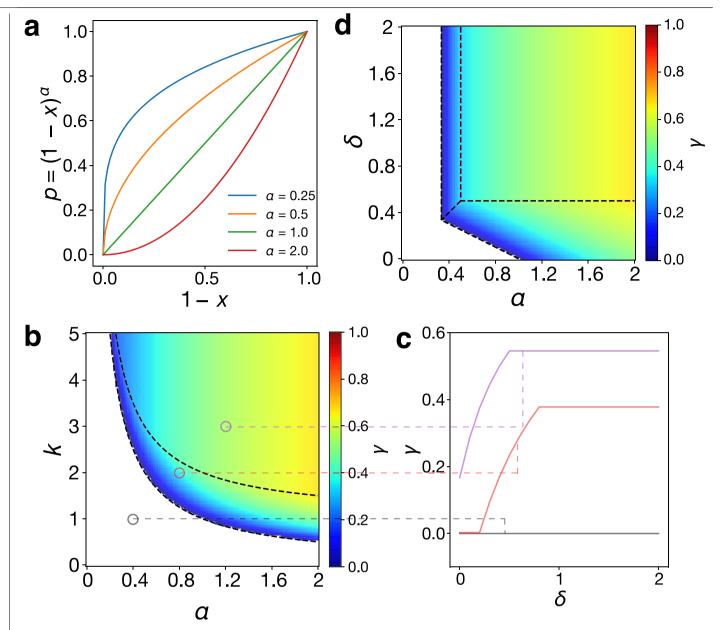
P=0.400, 0.859, 0.395), but quickly diverge for second failures (two-sided Welch's t-test;  $P=2.08\times10^{-3}$ ,  $6.70\times10^{-3}$ ,  $3.76\times10^{-3}$ ). The successful group also shows a significant improvement in performance (one-sided Welch's t-test;  $P=2.55\times10^{-2}$ ,  $5.65\times10^{-2}$ ,  $3.77\times10^{-2}$ ), which is absent for the unsuccessful group (one-sided Welch's t-test; P=0.970, 0.901, 0.967). Data are mean  $\pm$  s.e.m.  $\mathbf{j}$ - $\mathbf{l}$ , AUC score of predicting ultimate success as we focus on all samples ( $\mathbf{j}$ ), samples of human-targeted attacks ( $\mathbf{k}$ ) and include vague data on fatalities ( $\mathbf{l}$ ). The centres and error bars of AUC scores denote the mean  $\pm$  s.e.m calculated from tenfold cross-validation over 50 randomized iterations.  $\mathbf{m}$ - $\mathbf{o}$ , Temporal scaling patterns as we change the threshold for the successful group to fatal attacks that killed at least 5 ( $\mathbf{m}$ ), 10 ( $\mathbf{n}$ ) and 100 ( $\mathbf{o}$ ) people. \*P<0.1, \*\*P<0.05, \*\*\*P<0.01, NS, P<20.1.



**Extended Data Fig. 9** | See next page for caption.

Extended Data Fig. 9 | Additional robustness checks. a-i, Robustness check as we control for temporal variation. **a**-**c**, Failure streak in science (**a**), entrepreneurship (b) and security (c). Blue circles represent real data of successful groups and dashed lines represent fitted Weibull distributions. **d-f**, Temporal scaling patterns in science (**d**), entrepreneurship (**e**) and security (f). The shaded area shows mean  $\pm$  s.e.m. of  $T_n$  (log scale).  $\mathbf{g}$ - $\mathbf{i}$ , Performance dynamics in science ( $\mathbf{g}$ , n = 628, 145, 571, 123, from left to right), entrepreneurship ( $\mathbf{h}$ , n = 248, 1,332, 237, 1,312, from left to right) and security (i, n = 231, 173, 229, 174, from left to right). The successful and unsuccessful groups that experienced a large number of consecutive failures before their last attempt (at least 5 for  $D_1$ , 3 for  $D_2$  and 2 for  $D_3$ ) appear indistinguishable for first failures (two-sided weighted Welch's t-test; P = 0.814, 0.728, 0.330) but quickly diverge for second failures (two-sided weighted Welch's t-test;  $P=1.80\times10^{-2}$ ,  $3.10 \times 10^{-2}$ ,  $4.56 \times 10^{-2}$ ). The successful group also shows significant improvement in performance (one-sided weighted Welch's t-test;  $P = 2.10 \times 10^{-2}$ ,  $1.92 \times 10^{-2}$ ,  $4.53 \times 10^{-2}$ ), which is absent for the unsuccessful group (one-sided weighted Welch's *t*-test; P = 0.755, 0.175, 0.903). Data are mean  $\pm$  s.e.m.  $\mathbf{j} - \mathbf{l}$ , Performance dynamics as we compare first and halfway attempts in science (i. n = 628, 145, 582, 111, from left to right), entrepreneurship (**k**, n = 248, 1, 332, 240, 1,294, from left to right) and security ( $\mathbf{l}$ , n = 231, 173, 228, 175, from left to right). The successful and unsuccessful groups that experienced a large number of consecutive failures before their last attempt (at least 5 for  $D_1$ , 3 for  $D_2$  and 2 for  $D_3$ ) appear indistinguishable for first failures (two-sided Welch's t-test; P = 0.898, 0.671, 0.289) but diverge for halfway failures (two-sided Welch's t-test;  $P=2.18\times10^{-5}$ ,  $1.34\times10^{-2}$ ,  $1.34\times10^{-2}$ ). The successful group also shows significant

improvement in performance (one-sided Welch's t-test;  $P = 2.35 \times 10^{-2}$ ,  $4.54 \times 10^{-2}$ ,  $3.69 \times 10^{-2}$ ), which is absent for the unsuccessful group (one-sided Welch's t-test; P = 0.992, 0.252, 0.955). Data are mean  $\pm$  s.e.m.  $\mathbf{m} - \mathbf{o}$ , Performance dynamics as we compare the first and penultimate attempts in science (m, n = 628, 145, 896, 87, from left to right), entrepreneurship (**n**, n = 248, 1, 332, 227, 1,199, from left to right) and security ( $\mathbf{o}$ , n = 231,173,230,173, from left to right). The successful and unsuccessful groups that experienced a large number of consecutive failures before the last attempt (at least 5 for  $D_1$ , 3 for  $D_2$  and 2 for  $D_3$ ) appear indistinguishable for first failures (two-sided Welch's t-test, P = 0.898, 0.671, 0.289) but diverge for penultimate failures (two-sided Welch's t-test;  $P = 8.50 \times 10^{-8}$ ,  $3.12 \times 10^{-2}$ ,  $1.13 \times 10^{-2}$ ). The successful group also shows a significant improvement in performance (one-sided Welch's t-test;  $P = 5.79 \times 10^{-9}$ ,  $4.30 \times 10^{-2}$ ,  $1.33 \times 10^{-2}$ ), which is absent for the unsuccessful group (one-sided Welch's t-test; P = 0.980, 0.138, 0.923). Data are mean  $\pm$  s.e.m. p-r, The correlation between length of failure streak and initial performance (samples with repeated failures) in science ( $\mathbf{p}$ , n = 12,171), entrepreneurship ( $\mathbf{q}$ , n=2,086) and security ( $\mathbf{r}$ , n=441). Correlation is weak across all three datasets (Pearson correlation; r = -0.051, -0.011, -0.107 for  $\mathbf{p}$ ,  $\mathbf{q}$ ,  $\mathbf{r}$ , respectively).  $\mathbf{s} - \mathbf{u}$ , Length of failure streak still follow fat-tailed distributions conditional on bottom 10% initial performance samples in science ( $\mathbf{s}$ , n = 6,339), entrepreneurship ( $\mathbf{t}$ , n = 2,438) and security (**u**, n = 1,092). Two-sided Kolmogorov-Smirnov test between sample and exponential distributions rejects the hypothesis that the two distributions are identical with *P* < 0.01. \**P* < 0.1, \*\**P* < 0.05, \*\*\**P* < 0.01, NS, P≥ 0.1.



**Extended Data Fig. 10** | **Generalization of the** k **model. a**, The  $\alpha$  parameter connects the potential to improve (1-x) with the likelihood of creating new versions p through  $p = (1-x)^{\alpha}$ . **b**, Phase diagram of the  $k-\alpha$  model. The two-dimensional parameter space is separated into three regimes, with boundaries at  $k\alpha = 1$  and  $(k-1)\alpha = 1$ . **c**, The impact of  $\delta$  parameter on scaling exponent y for

given k=1,2,3 and  $\alpha=0.4,0.8,1.2$ . We find that  $\delta$  may affect the temporal scaling parameter when it is small, but has no further effect beyond a certain point  $\delta^*=\min(\alpha,1/(k-1))$ . **d**, Phase diagram of the  $k-\alpha-\delta$  model for k=3, with boundaries at  $\alpha=\delta$ ,  $(k-1)\delta=1$ ,  $(k-1)\delta+\alpha=1$ ,  $k\alpha=1$  and  $(k-1)\alpha=1$ , respectively.



Corresponding author(s):	Dashun Wang
Last updated by author(s):	Sep 2, 2019

# **Reporting Summary**

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics				
For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.				
n/a Confirmed				
The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement				
A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly				
The statistical test(s) used AND whether they are one- or two-sided Only common tests should be described solely by name; describe more complex techniques in the Methods section.				
A description of all covariates tested				
A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons				
A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)				
For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i> ) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted Give <i>P</i> values as exact values whenever suitable.				
For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings				
For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes				
$\square$ Estimates of effect sizes (e.g. Cohen's $d$ , Pearson's $r$ ), indicating how they were calculated				
Our web collection on <u>statistics for biologists</u> contains articles on many of the points above.				
Software and code				
Policy information about availability of computer code				
Data collection  This paper makes use of restricted access data from the National Institutes of Health, protected by the Privacy Act of 1974 as amended (U.S.C. 552a). De-identified data necessary to reproduce all plots and statistical analyses will be made freely available. Those wishing to access the raw data may apply for access following the procedures outlined in the NIH Data Access Policy document available at http://report.nih.gov/pdf/DataAccessPolicy.pdf. The VentureXpert database is available via Thomson Reuters. The Global Terrorism Database is publicly available at https://www.start.umd.edu/gtd/.				
Data analysis Data analyses were conducted using Python 3.4. Regression analysis were conducted using Stata 14.				
For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewer We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.				

#### Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

De-identified data necessary to replicate results of this study will be made freely available.

Field-specific reporting
Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Behavioural & social sciences For a reference copy of the document with all sections, see <u>nature.com/documents/nr-reporting-summary-flat.pdf</u>

Life sciences

Behavioural	&	social	sciences	study	design
				/	

	, 0
All studies must disclo	ose on these points even when the disclosure is negative.
Study description	A quantitative study of dynamics of failure based on pre-existing datasets.
Research sample	We collected three large-scale datasets from three domains: (1) R01 grant applications ever submitted to the National Institutes of Health (NIH), (776,721 applications by 139,091 investigators from 1985 to 2015); (2) Start-up investment records from VentureXpert database (58,111 startup companies involving 253,579 innovators); and (3) Terrorist attack data from Global Terrorism Database (70,350 terrorist attacks by 3,178 terrorist organizations from 1970 to 2017).
Sampling strategy	No statistical methods were used to predetermine sample size.
Data collection	This study is based on pre-existing datasets.
Timing	The NIH dataset was collected in 2016. The VentureXpert and GTD datasets were collected in 2017.
Data exclusions	The analysis has no data exclusions. Selection criteria within a dataset are described in the supplementary information.
Non-participation	There are no participants in this study.
Randomization	This is a data driven study, not a randomized experiment.

Ecological, evolutionary & environmental sciences

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems		Me	Methods	
n/a	Involved in the study	n/a	Involved in the study	
$\boxtimes$	Antibodies	$\boxtimes$	ChIP-seq	
$\boxtimes$	Eukaryotic cell lines	$\bowtie$	Flow cytometry	
$\boxtimes$	Palaeontology	$\bowtie$	MRI-based neuroimaging	
$\boxtimes$	Animals and other organisms		•	
$\boxtimes$	Human research participants			
$\boxtimes$	Clinical data			

# Hierarchical organization of cortical and thalamic connectivity

https://doi.org/10.1038/s41586-019-1716-z

Received: 16 April 2018

Accepted: 24 September 2019

Published online: 30 October 2019

Julie A. Harris<sup>1,5</sup>\*, Stefan Mihalas<sup>1,5</sup>, Karla E. Hirokawa<sup>1</sup>, Jennifer D. Whitesell<sup>1</sup>, Hannah Choi<sup>1,2</sup>, Amy Bernard<sup>1</sup>, Phillip Bohn<sup>1</sup>, Shiella Caldejon<sup>1</sup>, Linzy Casal<sup>1</sup>, Andrew Cho<sup>1</sup>, Aaron Feiner<sup>1</sup>, David Feng<sup>1</sup>, Nathalie Gaudreault<sup>1</sup>, Charles R. Gerfen<sup>3</sup>, Nile Graddis<sup>1</sup>, Peter A. Groblewski<sup>1</sup>, Alex M. Henry<sup>1</sup>, Anh Ho<sup>1</sup>, Robert Howard<sup>1</sup>, Joseph E. Knox<sup>1</sup>, Leonard Kuan<sup>1</sup>, Xiuli Kuang<sup>4</sup>, Jerome Lecoq<sup>1</sup>. Phil Lesnar<sup>1</sup>. Yaoyao Li<sup>4</sup>. Jennifer Luviano<sup>1</sup>. Stephen McConoughev<sup>1</sup>. Marty T. Mortrud<sup>1</sup>, Maitham Naeemi<sup>1</sup>, Lydia Ng<sup>1</sup>, Seung Wook Oh<sup>1</sup>, Benjamin Ouellette<sup>1</sup>, Elise Shen<sup>1</sup>, Staci A. Sorensen<sup>1</sup>, Wayne Wakeman<sup>1</sup>, Quanxin Wang<sup>1</sup>, Yun Wang<sup>1</sup>, Ali Williford<sup>1</sup>, John W. Phillips<sup>1</sup>, Allan R. Jones<sup>1</sup>, Christof Koch<sup>1</sup> & Hongkui Zeng<sup>1</sup>

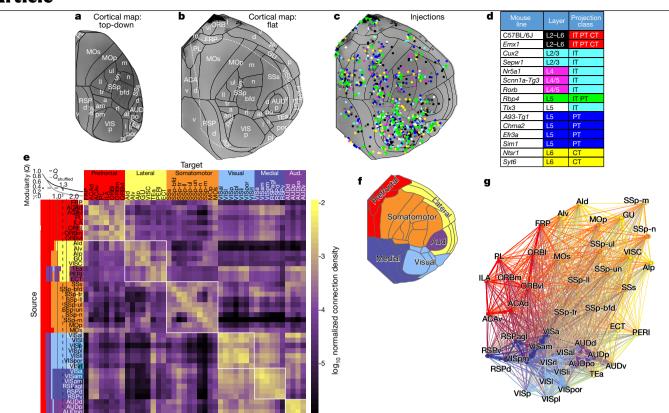
The mammalian cortex is a laminar structure containing many areas and cell types that are densely interconnected in complex ways, and for which generalizable principles of organization remain mostly unknown. Here we describe a major expansion of the Allen Mouse Brain Connectivity Atlas resource<sup>1</sup>, involving around a thousand new tracer experiments in the cortex and its main satellite structure, the thalamus. We used Cre driver lines (mice expressing Cre recombinase) to comprehensively and selectively label brain-wide connections by laver and class of projection neuron. Through observations of axon termination patterns, we have derived a set of generalized anatomical rules to describe corticocortical, thalamocortical and corticothalamic projections. We have built a model to assign connection patterns between areas as either feedforward or feedback, and generated testable predictions of hierarchical positions for individual cortical and thalamic areas and for cortical network modules. Our results show that cell-class-specific connections are organized in a shallow hierarchy within the mouse corticothalamic network.

Cognitive processes and voluntary control of behaviour originate in the cortex. To understand how incoming sensory information is processed and integrated with past experiences and current states in order to generate appropriate behaviour requires knowledge of the anatomical patterns and rules of connectivity between cortical areas. Connectomes—complete descriptions of brain wiring<sup>2</sup>—exist at different levels of spatial granularity, from single cells to populations of cells and entire areas (micro-, meso-, and macro-scale). Common organizational features of macro- and meso-scale cortical connectivity have been distilled across data sets<sup>1,3-7</sup>, often using graph theory approaches to describe network architecture<sup>8</sup>. For example, cortical areas have unique patterns of connections ('fingerprints'), connection strengths follow a log-normal distribution spanning more than four orders of magnitude<sup>1,4</sup>, and the organization of cortical areas is modular, with distinct modules corresponding to specific functions<sup>3,9-11</sup>.

The concept of hierarchical organization 12,13 is important for understanding the cortex, and has inspired the development of neural network methods in deep machine learning<sup>14</sup>. A hierarchy of cortical areas was first derived by mapping anatomical patterns of corticocortical (CC) connections onto feedforward and feedback directions. In primate, feedforward connections were characterized by dense axon terminations in layer (L)4 of the target area, and feedback connections as dense terminals in superficial and deep layers (avoiding L4)<sup>12,15</sup>. Differences in the layers of origin are also associated with feedforward and feedback connections 12,16. It is still unclear whether the concept of a cortical hierarchy, which has been derived largely from sensory systems, can be applied globally across the entire cortex, and how it arises from connections made by different classes of neurons. Each cortical region comprises distinct types of excitatory neurons that are largely organized by layers, but also by long-distance projection patterns: intratelencephalic (IT) in L2–L6, pyramidal tract (PT) in L5, and corticothalamic (CT) in  $L6^{17,18}$ .

Thalamic nuclei make important contributions to cortical function. They serve as a 'relay' for primary sensory information, and are well positioned to influence cortical information processing through reciprocal or transthalamic loops 19,20. Thalamocortical (TC) projection neurons are classified into three major classes<sup>21-23</sup>: core, intralaminar, and matrix. Like CC projections, feedforward and feedback rules have been proposed for TC and CT projections. Core projections (to L4) are described as 'driver' (feedforward) and matrix projections (to L1) as 'modulator' (feedback)<sup>19</sup>. For CT connections, input from L6 is considered feedback, and from L5 feedforward<sup>20</sup>.

We hypothesize that a unifying hierarchical organization across the entire cortex and its major input structure, the thalamus, is governed by a set of anatomical rules for CC, CT and TC connections. By using diverse Cre driver mouse lines to selectively label cells from different cortical layers and classes<sup>24-27</sup>, we have substantially expanded the



**Fig. 1**| **Cortical tracer experiments and network modularity. a**, Top-down view of the right cortical hemisphere in CCFv3. **b**, A virtual cortical flat map shows all 43 annotated areas. The white dotted line indicates the boundaries of what is visible in **a**. **c**, Cortical injection locations plotted on the flat map. **d**, Key summarizes layer and projection class selectively for 15 mouse lines. The colour code is also used in **c**; experiments in lines not listed are coloured dark grey. **e**, Matrix shows ipsilateral normalized connection densities between 43 cortical areas. Top left corner: the modularity metric (Q) and  $Q_{\text{shuffled}}$  are plotted for each y level. Colours to the left of each row indicate community structure at

 $\gamma$  = 0–2.5. Community structure was determined independently for each value of  $\gamma$ , but colours were matched to show how communities split as  $\gamma$  increases. Columns are coloured by the six modules identified at  $\gamma$  = 1.3. **f**, Cortical regions on the flat map colour-coded by module affiliation at  $\gamma$  = 1.3. **g**, Network diagram shows ipsilateral CC connections using a force-directed layout algorithm. Nodes are colour-coded by module. Edge thickness shows relative normalized connection density. Edges between modules are coloured as a blend of the connected node colours. For all abbreviations of structure names, see Supplementary Table 3.

Allen Mouse Brain Connectivity Atlas resource (http://connectivity. brain-map.org¹), adding 1,256 new tracing experiments. Our findings follow analyses of projection patterns spanning nearly the entire mouse cortex and thalamus, and show how these patterns relate to layer and cell class. We test the above hypothesis by building a computational hierarchical model using anatomical rules derived from observations of axon termination patterns. Our results show that the mouse cortex and thalamus form an integrated hierarchical organization.

#### Cre drivers for cortical projection mapping

Our goal for expanding the Allen Mouse Brain Connectivity Atlas¹ was to create a map of all interareal projections that originate from neurons of different cell classes within a given source. Here, we used 50 mouse lines (wild-type C57BL/6J mice and 49 Cre driver lines) for cortical projection mapping. We injected Cre-dependent enhanced green fluorescent protein (EGFP) or synaptophysin–EGFP viral tracers to selectively trace axons from Cre⁺ neurons (see Extended Data Fig. 1a–c for virus comparison). Using our high-throughput imaging and informatics pipeline approach, we produced 1,081 cortical tracer experiments suitable for analyses (see Methods and Supplementary Tables 1, 2). To visualize coverage, we plotted injection locations for all experiments on a cortical surface flat map of the 3D Allen Common Coordinate Framework reference atlas (CCFv3, Fig. 1a–c). High-resolution image series, visualization tools, and quantification of injection sites and brain-wide targets are accessible through our data portal (http://connectivity.brain-map.org).

We inspected brain-wide axonal projection patterns to manually classify each experiment into one of six layer and projection classes: (1) IT PT CT: labelled axons originate from all source layers and terminate in all target regions (ipsilateral and contralateral cortex and striatum, thalamus, and midbrain, pons and/or medulla); (2) IT PT: labelled axons observed in all target regions, but the injection site did not include L6 neurons; (3) IT: labelled axons restricted to ipsilateral and contralateral cortex and striatum; (4) PT: labelled axons projected ipsilaterally and subcortically; (5) CT: labelled axons project almost exclusively to thalamus from L6; and (6) local: no (or few) long-distance axons present (Extended Data Fig. 2a, Supplementary Table 1). Manual assignment to projection class was consistent with the results of unsupervised clustering (Extended Data Fig. 2b–d) and previous characterizations<sup>25</sup>. We also characterized layer selectivity in the source for each Cre line on the basis of injection sites (Extended Data Fig. 2e).

From these data, we chose a core set of 15 lines that we used to comprehensively map connectivity from different projection neuron classes across cortical layers (Fig. 1d), resulting in 849 experiments used for subsequent analyses of CC and CT projections. We did not identify a suitable Cre line for L6 IT $^{28}$ .

#### **Corticocortical connectivity modules**

Previous network analyses revealed a modular community structure in the mouse brain, including in the isocortex<sup>9</sup>. To determine whether our data set demonstrated a similar network architecture, we constructed

an ipsilateral cortical connectivity matrix (Fig. 1e) using a data-driven model based on wild-type mice<sup>29</sup>. We analysed the network structure of this matrix using the Louvain algorithm<sup>30</sup>, which maximizes a modularity metric (Q) to identify groups of nodes (cortical areas) that are most densely connected to each other compared to a randomized network. To identify stable modules, we systematically varied the spatial resolution parameter, y, from 0 to 2.5, measured Q at each value, and compared the results to a shuffled network. The mouse cortex was modular  $(Q > Q_{\text{shuffled}})$  for every value of  $\gamma$  above 0.3. We chose to focus on the six modules identified at y = 1.3 (Q = 0.36), where the difference between Q and  $Q_{\text{shuffled}}$  was maximal (0.22  $\pm$  0.017, mean  $\pm$  s.d.). We named these six modules for the areas assigned to each: prefrontal, lateral, somatomotor, visual, medial, and auditory. Even with substantial community structure, intracortical connections are dense between modules (Fig. 1f). The Louvain algorithm parameterizes edge strength only, with no constraint for spatial arrangement of nodes, but there is a clear spatial component, in that neighbouring areas usually belong to the same module (Fig. 1g). We directly tested the degree to which spatial proximity affects modularity by fitting a power-law to the distance component of the ipsilateral connectivity matrix, and then analysing the resulting residual matrix using the Louvain algorithm (Extended Data Fig. 3). Although fewer modules were present after accounting for distance, regions within them were generally still anatomically adjacent.

#### Corticocortical projections by layer or class

To investigate the contributions of distinct cell classes within each area to CC projections, we compiled 43 groups of spatially matched experiments, each having a 'complete' membership roster representing all layer classes (L2/3 IT, L4 IT, L5 IT PT, L5 IT, L5 PT and L6 CT) plus a wildtype or Emx1 IT PT CT data set (note that Cre mouse lines are referred to by gene symbol; for example, *Emx1* is *Emx1-IRES*-Cre). Projection class was confirmed for each experiment. These 43 anchor groups, composed of 364 experiments, represent 25 out of 43 CCFv3 cortical areas (Fig. 2a-d, Supplementary Table 4). From any given source, CC projections labelled from these Cre lines had similar overall patterns, but Rbp4 (L5 IT PT) consistently appeared to show the most extensive projections (Fig. 2e). Intracortical projections were labelled from all layers (L2/3-L6). The identification of interareal projections from L4 was unexpected, given canonical circuit descriptions, but is not without recent precedent <sup>28,31</sup>. To confirm that IT projections could truly be attributed to L4 neurons, we reconstructed the complete dendritic and axonal morphology of 25 sparsely labelled neurons following wholebrain fluorescence micro-optical sectioning tomography (fMOST) imaging<sup>32</sup>. We identified three classes of L4 neuron using morphological  $criteria^{33}, and \, confirmed \, that \, many, \, but \, not \, all, \, individual \, L4 \, cells \, sent$ axons to other cortical areas (Extended Data Fig. 4).

To make quantitative comparisons across Cre lines, we first manually identified true positive and negative connections for each experiment in the anchor groups (43 ipsilateral and 43 contralateral targets in 364 experiments, giving 31,304 connections checked; Supplementary Table 4). We noted when a target contained only fibres of passage, and considered it a true negative. Using automated segmentation and registration to CCFv3, we generated a weighted connectivity matrix (using normalized projection volume (NPV); see Methods) for each Cre line (wild-type and Emx1 were merged; Extended Data Fig. 1e-g), and applied the true positive mask to remove true negative connections (Fig. 2f). We selected only one anchor group per cortical region for visualization if there was a significant, positive correlation between *Rbp4* replicates (Spearman r > 0.8), resulting in 27 groups; 25 unique areas and two locations in the secondary motor area and supplemental somatosensory area.

Overall, the CC matrices revealed several features of layer- or class-specific connectivity in terms of the number and specificity of connections. The average 'out-degree' (number of targets; Fig. 2g) from Rbp4 was larger in both hemispheres than from any other Cre line. except for Tlx3 on the contralateral side. L5PT and L6CT lines had the fewest targets in both hemispheres, followed by the L2/3, L4, and L5 IT lines. For every line, there were fewer (or no) contralateral compared to ipsilateral connections.

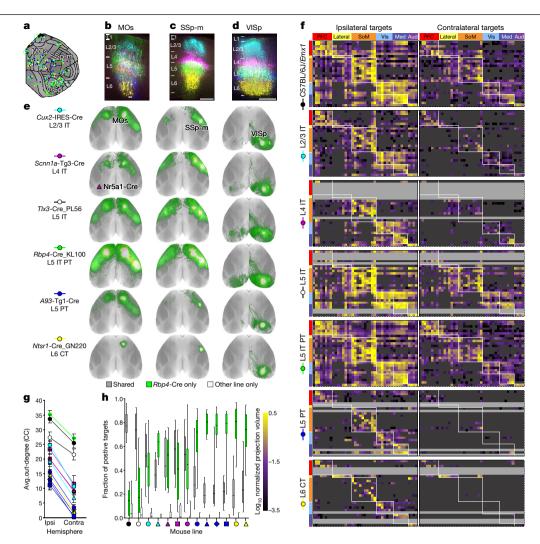
We measured the amount of overlap between the specific set of cortical targets contacted in each experiment and its Rbp4 anchor (Fig. 2h). Wild-type/Emx1 projections went to about 80% of the same targets as Rbp4 axons. A roughly equal number of targets was unique to either wild-type or *Emx1* (12.7 and 7%, respectively), perhaps owing to injection variability or differences in viral tracers. For every other Cre line, essentially all projections went to a subset of L5 Rbp4 targets (Fig. 2h, fewer than 5% of targets unique to any line). Within L5. IT cells had the most overlap with Rbp4 targets, whereas PT cells had the least (Fig. 2e, f). Most of the differences between L2/3 and L5 resulted from the presence of fewer projections to the contralateral hemisphere from L2/3 (Fig. 2g).

#### Thalamocortical projections by source or class

To investigate TC connections, we selected 81 out of 254 injection experiments in wild-type and Cre driver lines on the basis of the extent of anatomical restriction to a single region. These experiments covered 29 out of 44 thalamic nuclei in CCFv3 (Supplementary Tables 1, 2, Extended Data Fig. 5a). Most thalamic nuclei that are known to contain cortically projecting neurons were included, except for the posterior triangular thalamic nucleus (PoT), suprageniculate nucleus (SGN), anterodorsal nucleus (AD), and interanteromedial nucleus of the thalamus (IAM)<sup>23</sup>.

We visually inspected the brain-wide axonal projection patterns and classified these 81 experiments using previous definitions for core, matrix and intralaminar TC projection classes 18,23. Each experiment was manually assigned to one of four groups, or 'none' if no TC axons were observed (Fig. 3a): (1) core: labelled axons were observed in a small number of cortical targets with axons predominantly ramifying in L4; (2) intralaminar: labelled axons were predominantly observed in the striatum, with weak or diffuse cortical axons present; (3) matrix (focal): labelled axons targeted L1 in a small number of nearby targets; and (4) matrix (multiareal): labelled axons targeted L1 in a more distributed set of targets. Most thalamic nuclei could be assigned to one class, although this does not preclude regions having mixed classes (Fig. 3b, Supplementary Table 1). Only three regions (all primary sen $sorv\,thalamic\,nuclei)\,were\,assigned\,core-type\,projections-the\,ventral$ posterolateral nucleus (VPL), ventral posteromedial nucleus (VPM), and dorsal part of the lateral geniculate complex (LGd)<sup>19,21</sup>. Most thalamic sources were matrix or intralaminar.

Unlike the cortex, which is organized into distinct projection classes within layers of a single region, thalamic nuclei contain relatively homogenous populations of cortically projecting neurons<sup>34</sup>. As we used multiple Cre lines and wild-type mice for thalamic injections (Supplementary Table 1), we generated a TC connectivity matrix to compare the patterns of individual experiments (Fig. 3f). We manually identified true positive and true negative (including fibres of passage) connections for cortical targets (43 ipsilateral and 43 contralateral targets in 81 experiments, giving 6,966 connections manually checked; Supplementary Table 5), and performed hierarchical clustering on the masked weights (Fig. 3c). Most sources with multiple injections clustered together, even those from different lines. Exceptions included the mediodorsal nucleus (MD), where further analyses showed that precise location mattered more than Cre line (MD-1 experiments are in mid-to-caudal MD, MD-2 experiments are in the rostral portion). The specific patterns of cortical areas targeted by each cluster of thalamic nuclei were remarkably like the cortical modules defined by CC connections (Extended Data Fig. 5b).



**Fig. 2** | **Corticocortical projection patterns by layer and class. a**, Forty-three groups of experiments, spatially matched to one Rbp4 anchor (green dots). Most group members were less than 500 μm from the anchor (median, 296 μm). Green circles indicate the variance in distance to Rbp4 for each group. **b**-**e**, Data from three groups. **b**-**d**, Serial two-photon tomography (STPT) images at the centre of each injection site per Cre line were manually overlaid by finding the best match between the pial surface (top) and white matter boundary (bottom), then pseudocoloured by line. Scale bar, 250 μm. **e**, Top-down views of CC projections for spatially matched experiments. **f**, Directed, weighted connectivity matrices ( $27 \times 86$ ) for seven mouse lines: wild-type and the six Cre lines in **e**. Each row contains the  $\log_{10}$ -transformed NPVs from a single experiment in one of 27 source areas. Columns show cortical target

regions. Rows and columns follow the same order in each matrix. White boxes highlight regions in the same module. True negatives and passing fibres were masked out (dark grey). Rows for which an experiment was missing (often because of low Cre expression) are light grey. The colour map ranges from  $10^{-3.5}$  to  $10^{0.5}$  log NPV. It is truncated at both ends.  ${\bf g}$ , Mean out-degrees ( $\pm$  s.e.m.) across all sources for each Cre line plotted for ipsilateral and contralateral cortex.  ${\bf h}$ , The fraction of true positive targets shared by each line with its Rbp4 anchor is shown in the box plot (grey). The fraction of positive targets unique to Rbp4 (green) or to the line indicated (white) are also shown. Box plots show median and interquartile range (IQR). Whiskers show minimum and maximum values.

#### Corticothalamic projections by layer or class

We also used the Rbp4 anchored cortical experiments to generate weighted CT connectivity matrices. We again identified true positive and true negative connections, this time for all thalamic targets (44 ipsilateral and 44 contralateral targets in 256 experiments, giving 22,528 CT connections manually checked; Supplementary Table 6). As expected from previous publications, most cortical projections to the thalamus were observed in L5 PT and L6 CT Cre lines (and wild-type mice), with minimal to no true positive connections from L2/3, L4, and L5 IT lines (Supplementary Table 6). We focused our subsequent analyses on Rbp4 to represent L5 PT, given its more comprehensive coverage. Connection strengths were significantly correlated between the L6 CT lines Ntsr1 and Syt6 (P < 0.0001, Fig. 4d), so we averaged or merged these data (Fig. 4b).

The L5 and L6 CT matrices appear similar, but have quantitative differences (Fig. 4a, b). Many thalamic targets receive inputs from both

layers (Fig. 4c), and the connection weights for shared targets is significantly correlated (P<0.0001, Fig. 4f). However, this coefficient (0.65) is smaller than that between replicate experiments (0.84, Fig. 4e) and between the L6 lines (0.77, Fig. 4d). We calculated and visualized relative differences in input strength from L5 and L6 for every source—target pair in the anchor group matrix (Fig. 4g). Some targets are contacted more, or less, by L5 or L6 depending on source region, but other targets have stronger L5 or L6 input regardless of source (bands of a single colour down a column in Fig. 4g).

Notably, some CT projections clearly travel through thalamic regions before reaching their final targets, but also form synapses in those areas. Although entire regions containing only passing fibres were masked out, the remaining connections can contain a mix of fibres and terminals. To determine the effect of this kind of axonal trajectory on quantification of CT connection strengths, we compared a subset of spatially matched data sets in L5 and L6 Cre lines using

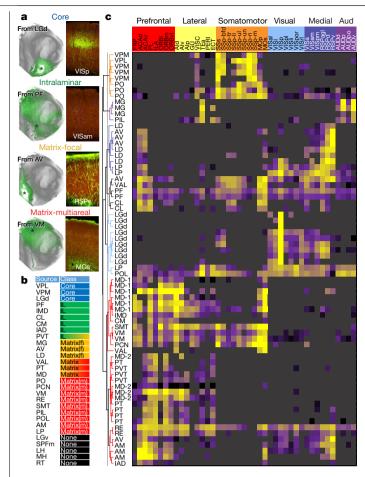


Fig. 3 | Thalamocortical projection patterns by region and class. a, Left, flatmap views show TC projections labelled from the region indicated. Right, STPT images from the centre of a cortical target (asterisk on left) show example axon  $lamination\ patterns\ associated\ with\ each\ projection\ class.\ \textbf{b},\ Key\ summarizes$ projection class assigned for 29 thalamic nuclei. c, The TC connectivity matrix  $(70 \times 43)$  for individual viral tracer injection experiments with verified cortical projections. Each row shows log<sub>10</sub>-transformed NPVs from one experiment to the 43 ipsilateral cortical targets (columns). Cre line names for each row are in Supplementary Table 5). Unsupervised hierarchical clustering, using Spearman correlation and average linkages, revealed seven clusters containing thalamic regions with cortical projection patterns resembling the cortical modules. Matrix colour map as in Fig. 2.

synaptophysin-EGFP to preferentially label terminals over fibres (Extended Data Fig. 6; see Methods). We found a strong linear relationship between measured connection strengths, and between relative differences between L5 and L6, with these two tracers, showing that EGFP tracer results can be used confidently for quantitative estimates of CT strengths. Nevertheless, this is an important consideration, as across the entire brain connection strengths from synaptophysin-EGFP experiments are on average lower than when using EGFP (Extended Data Fig. 1c), specifically by about 0.5 log units for Rbp4 CT targets (Extended Data Fig. 6k).

#### Laminar termination patterns in cortex

Using automated image registration to CCFv3, we quantified projection strengths by layer within each cortical target (registration precision in Extended Data Fig. 7a-c, Supplementary Table 7; see Methods). Then, to identify common laminar termination patterns across all sources and lines, we performed unsupervised hierarchical clustering with the complete data set (849 cortical and 81 thalamic experiments). Data had to pass three filters, as follows. (1) Target connection strength (log<sub>10</sub>

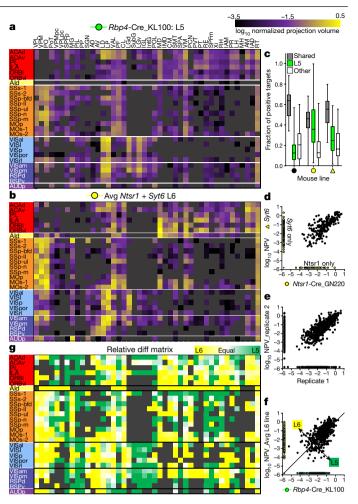
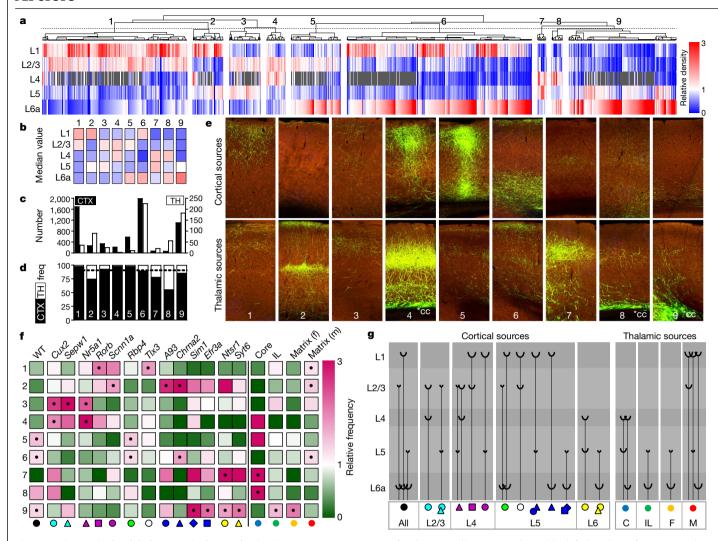


Fig. 4 | Corticothalamic projections from layers 5 and 6, a, b, CT connectivity matrices (27 × 44) for L5 (a, Rbp4) and L6 (b, average of Ntsr1 and Syt6). Each row shows log<sub>10</sub>-transformed NPVs from one of the 27 cortical source areas in Fig. 2 to the 44 ipsilateral thalamic target regions (columns). **c**, The fraction of true positive CT targets shared by wild-type (black circle) and each L6 line (yellow) with its Rbp4 anchor is plotted in the box plot (grey). The fraction of positive targets unique to Rbp4 (green) or unique to the L6 line (white) are also shown. Box plots show median and IQR. Whiskers show minimum and maximum values. **d**, log NPVs for thalamic targets shared by Ntsr1 and Syt6 were significantly correlated (Spearman r = 0.77, P < 0.0001). e, log NPVs for thalamic targets shared by replicate experiments in the same Cre line less than 500  $\mu m$ apart were significantly correlated (Spearman r = 0.84, P < 0.0001). **f**, The average log NPVs originating from L6 are plotted against L5 for all spatially matched experiments (Spearman r = 0.65, P < 0.0001). **g**, The matrix shows the relative difference for each source × target connection originating from L5 versus L6 (L5 - L6/L5 + L6).

transformed NPV) was above -1.5. This threshold was chosen after analysing NPV frequency distributions for a set of manually verified true positive and true negative connections (Extended Data Fig. 7d). (2) Percentage of infection volume in the primary source was more than 50%. (3) Self-to-self (within area) projections were removed. Following these steps, if present, multiple experiments were averaged, resulting in a total of 7,063 (660 thalamus, 6,403 cortex) unique source-line-target connections (Supplementary Table 8). We identified nine clusters (Fig. 5a). The median relative density values for each layer and the overall frequencies of these clusters are shown in Fig. 5b-d. Representative images from specific connections (a given source-line-target) assigned to each cluster from a cortical and thalamic source are shown in Fig. 5e.

A summary of cluster representation shows that each cortical Cre line and TC projection class is associated with more than one type of



 $Fig.\,5\,|\,Corticocortical\,and\,thal a mocortical\,target\,lamination\,patterns.$ 

**a**, Unsupervised hierarchical clustering on relative projection density per layer. Each column is a unique combination of mouse line, cortical or thalamic source area, and cortical target. Connections to agranular (no L4) regions are coloured grey for L4. The dotted line indicates where the dendrogram was cut into nine clusters. **b**, Median relative density by layer for each cluster. **c**, Number of cortical or thalamic connections in each cluster, plotted on the left and right *y*-axis, respectively. **d**, The frequency of cortical and thalamic targets assigned to each cluster. The dotted line indicates the overall frequency of CC targets in the entire data set (90.53%). **e**, Representative STPT images show axonal lamination patterns from a connection assigned to each cluster from cortex or thalamus. In columns 4, 8, and 9, thalamic axons passing through the

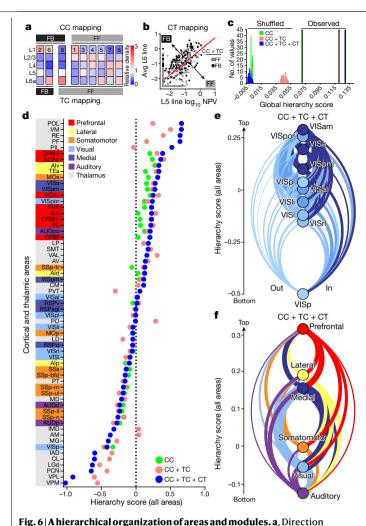
superficial corpus callosum are indicated (\*cc). **f**, The relative frequency with which each cortical Cre line and TC projection class appears in the clusters. The fraction of experiments in a cluster belonging to each Cre line or class was divided by the overall frequency of experiments from that Cre line or class. A relative frequency value of 1 (white) indicates that the Cre line appeared in that cluster with the same frequency as in the entire data set. Values below 1 (green) indicate lower and above 1 (pink) indicate higher than expected frequency in a cluster. Dots indicate significant positive enrichment in that cluster (Fisher's exact test, P < 0.0001). **g**, Schematic diagram shows significantly enriched axon lamination patterns associated with each layer and/or class of origin in the source area.

target layer pattern (Fig. 5f). The most common, and significantly enriched, laminar patterns from each are schematized in Fig. 5g (Fisher's exact test, P < 0.0001). L2/3 and L4 (Nr5a1) neurons project predominantly to middle layers (L2/3, L4, and L5), avoiding L1. Other L4 neurons project to L1 and either L2/3 or L5, avoiding L4 and L6. In L5, when both IT and PT classes are labelled, as in the Rbp4 line, projections target L6 and either L1 or L2/3. L5 IT neurons predominantly target superficial layers (L1 and L2/3). L5 PT neurons target either deep layers only (L5 and L6) or deep layers and L1, consistent with the L5 Rbp4 patterns representing both IT and PT patterns. L6 CT neurons project predominantly to deep layers. From thalamic sources, core neurons project to L4 and either L5 or L6, intralaminar and matrix-focal neurons preferentially project to L5 and L6, and connections coming from matrix-multiareal sources all project to L1, with differing proportions in other layers.

#### Hierarchy of cortical and thalamic areas

We hypothesized that the above anatomical rules could be used across all cortical and thalamic regions to build a testable hierarchical model to predict the direction of information flow. We used cortical Cre line experiments (Fig. 6) because they allow incorporation of the specific layer termination patterns related to cell class, but results are also provided using wild-type data (Extended Data Fig. 10).

We used an unbiased approach to identify the most optimal label for each of the nine clusters; feedforward (FF) or feedback (FB). We defined an initial hierarchical position for each cortical area (as both a source and target) using the averaged difference between FB and FF connections, normalized by a confidence measure for each cortical Cre line (Eqs. (2–4) in Methods; see also results without this confidence term in Extended Data Fig. 10). We searched over all possible mappings between the nine layer patterns and directional assignments, and



mapping results for CC and TC terminal layer patterns. Median relative density by layer for each cluster shown from Fig. 5b. b, Direction mapping for CT connections. Scatterplot shows the  $\log_{10}$ -transformed NPVs for every CT connection from L5 and L6. Points are colour-coded by the mapping (FF or FB) predicted from the CC + TC hierarchy. Red line, linear discriminant analysisassigned connections (below, FF; above, FB). c, Global hierarchy scores (Eq. (17)) for CC connections only (green), compared to the scores when TC and CT connections are included sequentially (pink, blue). Scores for the original, observed, data are shown as single outlined bars. Distributions of hierarchy scores were obtained from shuffled data sets (n = 100). The medians of the shuffled distributions estimate the lower bound (0.001, 0.044 and -0.002 for CC, CC + TC and CC + TC + CT, respectively). d, Thirty-seven cortical areas and twenty-four thalamic nuclei rank-ordered by their CC + TC + CT hierarchy scores. Scores for each area using only CC or CC + TC connections are also plotted. y-axis labels are colour-coded by module assignment for cortical areas. e, Network diagram showing interconnections of all cortical visual areas (light blue, visual module; dark blue, medial module). Edge width represents relative connection density from Fig. 1e. The curved lines show outputs (left) and inputs (right) to each node. Nodes are positioned along a single axis based on hierarchical score. f, Intermodule network diagram. Edge width represents sum of connection densities from Fig. 1e.

determined which mapping resulted in the most self-consistent initial hierarchy (maximized the CC global hierarchy score, which measures how consistent the obtained hierarchy is with directions of individual connections; Eq. (5) in Methods). For CC connections, clusters 2, 6, and 9 were assigned to one direction, and 1, 3, 4, 5, 7 and 8 to the other (Fig. 6a). For TC projections, clusters 2 and 6 were assigned to the same direction and the rest to the other (Eqs. (8-10) in Methods). Cluster 9 switched directions for CC and TC. We confidently labelled these

two directions as either FF or FB on the basis of extensive anatomical analyses relating our observed layer patterns to known hierarchical order and rules between reciprocally connected regions<sup>35–38</sup> (Extended Data Figs. 8, 9). After obtaining initial hierarchy positions with the most optimal mappings, scores were iterated to further refine the hierarchy (Eas. (6.7.11.12) in Methods).

To label CT connections, we used the Cre-defined L5 and L6 projection strengths (thresholded by log<sub>10</sub>-transformed NPV > -2.5; Extended Data Fig. 7e; Eq. (13) in Methods). Linear discriminant analysis was applied to assign CT connections into the FF or FB class that was most self-consistent with the direction predicted from a CT hierarchy constructed first from CC and TC projections (Fig. 6b, Supplementary Table 9; Eq. (14) in Methods).

Using these rules, we obtained three versions of hierarchy based on (1) CC connections only, (2) CC and TC connections, and (3) CC, TC, and CT connections. We demonstrate that there is significant hierarchical organization by comparing global hierarchy scores with corresponding distributions of scores from shuffled connections (Fig. 6c, see z-scores in Extended Data Fig. 10f). Adding thalamus connections essentially doubled the hierarchy scores (0.069, 0.120, and 0.128 for CC, CC + TC, and CC+TC+CT, respectively). Nonetheless, by comparing the global hierarchy scores with their maxima (0.679, 0.636, and 0.683; see Methods), it appears to be a rather shallow hierarchy.

The final hierarchical positions for 37 cortical areas and 24 thalamic nuclei are presented in Fig. 6d (scores in Supplementary Table 9). Most thalamic regions are located at the bottom or top, suggesting that they have pure driver or modulator effects on the cortical areas with which they are connected. Several thalamic nuclei appear mid-hierarchy, indicating more balanced numbers of FF and FB connections. For cortical regions, primary visual cortex is at the bottom and the prefrontal area ORBvl (ventrolateral part of the orbital area) is at the top. Predicted hierarchical positions were broadly similar across the three versions (CC, CC + TC, or CC + TC + CT). Most regions had only minor shifts in position. The largest shifts occurred in thalamic regions when adding CT connections. Hierarchies were also exceedingly robust to contributions from any single Cre line, or layer or projection class (Extended Data Fig. 10h).

We used similar methods to predict hierarchy for subsets of areas (visual cortex, Fig. 6e) and between modules (Fig. 6f). The intermodule hierarchy had a relatively low global score (0.07) compared to the allarea hierarchy (0.13), but it was more obviously organized into distinct levels: primary sensory modules at the bottom, lateral and medial modules in the middle, and prefrontal at the top.

#### Discussion

We used a genetic viral tracing approach, building on our previously established whole-brain imaging and informatics pipeline, to map projections originating from unique cell populations in the same cortical area, and from distinct projection classes in the thalamus. Our study represents a big step towards a true mesoscale connectome<sup>39</sup>. It will be informative for future connectome studies with more refined cell types and single cells<sup>40-42</sup>, which will no doubt reveal additional principles of cell-type-specific brain connectivity<sup>43</sup>. With these mesoscale data, we derived several generalizable anatomical rules of cortical and thalamic connections, and tested whether the organizing principle of a hierarchy applies to mouse cortex and thalamus.

The cortex is organized as a modular network<sup>3,9,11</sup>, which provides a structural view of possible paths of information flow, but does not impose direction or order onto that flow. By contrast, a hierarchy implies that interareal connections belong to at least two general types: feedforward or feedback. Specific anatomical projection patterns were previously associated with information transmission in these directions in primate and rodent visual cortex<sup>12,13,36,38</sup>. In our data, we observed many similar patterns. Two patterns that differed were the superficial

layer projections (cluster 1) and the deep layer projections (cluster 9). Felleman and van Essen<sup>12</sup> noted the occasional superficial-only pattern, but they called it feedback because it did not involve L4. Our results suggest this pattern is associated with feedforward. The strength and presence of projections between areas from the predominantly L4 Cre lines was also unexpected, given canonical circuit diagrams<sup>44</sup>, and might be explained by varying degrees of layer selectivity. However, by reconstructing the complete dendritic and axonal morphology of single cells, we directly show that L4 neurons, even spiny stellate cells, can in fact have long-range projections.

The hierarchy that we find is shallower than might have been expected, even with inclusion of thalamic regions. The difference between the lowest and the highest areas is less than two full levels, and the all-area hierarchy global score is at 19% between random and perfectly hierarchical. This might be characteristic of the mouse cortex, given its high connection density, particularly when considering all non-zero connection strengths<sup>45</sup>. We did not explicitly include strengths in computing hierarchy, except that weak connections were removed. Notably, hierarchical position alone does not explain all of the connections of a given area. This complexity may be why some have argued that the concept of a hierarchy is overly simplistic for describing functional properties<sup>46</sup>. Given the number of different connection types that arise from a single area, future computational models that incorporate more than feedforward and feedback labels will enable further insights into the organization of brain networks.

Cortical hierarchies were previously derived from classic anterograde or retrograde tracing without cell-class resolution. Using Cre lines, we have mapped both layer of origin and target lamination pattern in the same experiment. We found that L2/3 and L4 neurons have predominantly feedforward layer projection patterns, whereas L5 and L6 neurons have both feedforward and feedback patterns. However, these general relationships depend on the specific source–target connection and Cre line. The Cre data set, with all this detail, produced the most robust hierarchy (Extended Data Fig. 10f). However, our results from wild-type mice provide a solid benchmark for others interested in applying these hierarchical model algorithms to classic tracing data. The calculation of global hierarchy scores for other data sets will enable direct comparisons between species and quantitative assessments of how development or disease might affect hierarchical organization.

#### **Online content**

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1716-z.

- 1. Oh, S. W. et al. A mesoscale connectome of the mouse brain. *Nature* **508**, 207–214 (2014).
- Sporns, O., Tononi, G. & Kötter, R. The human connectome: a structural description of the human brain. PLOS Comput. Biol. 1, e42 (2005).
- 3. Zingg, B. et al. Neural networks of the mouse neocortex. Cell 156, 1096-1111 (2014).
- Markov, N. T. et al. A weighted and directed interareal connectivity matrix for macaque cerebral cortex. Cereb. Cortex 24, 17–36 (2014).
- Bota, M., Sporns, O. & Swanson, L. W. Architecture of the cerebral cortical association connectome underlying cognition. Proc. Natl Acad. Sci. USA 112, E2093–E2101 (2015).
- Scannell, J. W., Blakemore, C. & Young, M. P. Analysis of connectivity in the cat cerebral cortex. J. Neurosci. 15, 1463–1483 (1995).
- Swanson, L. W., Hahn, J. D. & Sporns, O. Organizing principles for the cerebral cortex network of commissural and association connections. *Proc. Natl Acad. Sci. USA* 114, E9692–E9701 (2017).
- Bullmore, E. & Sporns, O. Complex brain networks: graph theoretical analysis of structural and functional systems. Nat. Rev. Neurosci. 10, 186–198 (2009).
- Rubinov, M., Ypma, R. J. F., Watson, C. & Bullmore, E. T. Wiring cost and topological participation of the mouse brain connectome. *Proc. Natl Acad. Sci. USA* 112, 10032–10037 (2015).
- Wang, Q., Sporns, O. & Burkhalter, A. Network analysis of corticocortical connections reveals ventral and dorsal processing streams in mouse visual cortex. J. Neurosci. 32, 4386–4399 (2012).

- Swanson, L. W., Hahn, J. D., Jeub, L. G. S., Fortunato, S. & Sporns, O. Subsystem organization of axonal connections within and between the right and left cerebral cortex and cerebral nuclei (endbrain). Proc. Natl Acad. Sci. USA 115, E6910–E6919 (2018).
- Felleman, D. J. & Van Essen, D. C. Distributed hierarchical processing in the primate cerebral cortex. Cereb. Cortex 1, 1–47 (1991).
- Rockland, K. S. & Pandya, D. N. Laminar origins and terminations of cortical connections of the occipital lobe in the rhesus monkey. Brain Res. 179, 3–20 (1979).
- Riesenhuber, M. & Poggio, T. Hierarchical models of object recognition in cortex. Nat. Neurosci. 2, 1019–1025 (1999).
- Rockland, K. S. What do we know about laminar connectivity? Neuroimage 197, 772–784 (2019).
- Markov, N. T. et al. Anatomy of hierarchy: feedforward and feedback pathways in macaque visual cortex. J. Comp. Neurol. 522, 225–259 (2014).
- Shepherd, G. M. G. Corticostriatal connectivity and its role in disease. Nat. Rev. Neurosci. 14, 278–291 (2013).
- Harris, K. D. & Shepherd, G. M. G. The neocortical circuit: themes and variations. Nat. Neurosci. 18, 170–181 (2015).
- Sherman, S. M. Thalamus plays a central role in ongoing cortical functioning. Nat. Neurosci. 19, 533–541 (2016).
- Usrey, W. M. & Sherman, S. M. Corticofugal circuits: communication lines from the cortex to the rest of the brain. J. Comp. Neurol. 527, 640–650 (2019).
- 21. Jones, E. G. The Thalamus (Cambridge Univ. Press, 2007).
- 22. Jones, E. G. Viewpoint: the core and matrix of thalamic organization. *Neuroscience* **85**, 331–345 (1998)
- Clascá, F., Rubio-Garrido, P. & Jabaudon, D. Unveiling the diversity of thalamocortical neuron subtypes. Eur. J. Neurosci. 35, 1524–1532 (2012).
- Gong, S. et al. Targeting Cre recombinase to specific neuron populations with bacterial artificial chromosome constructs. J. Neurosci. 27, 9817–9823 (2007).
- Gerfen, C. R., Paletzki, R. & Heintz, N. GENSAT BAC cre-recombinase driver lines to study the functional organization of cerebral cortical and basal ganglia circuits. *Neuron* 80, 1368–1383 (2013).
- Harris, J. A. et al. Anatomical characterization of Cre driver mice for neural circuit mapping and manipulation. Front. Neural Circuits 8, 76 (2014).
- Daigle, T. L. et al. A suite of transgenic driver and reporter mouse lines with enhanced brain-cell-type targeting and functionality. Cell 174, 465–480.e22 (2018).
- Tasic, B. et al. Shared and distinct transcriptomic cell types across neocortical areas. Nature 563, 72–78 (2018).
- Knox, J. E. et al. High-resolution data-driven model of the mouse connectome. Netw. Neurosci. 3, 217-236 (2018).
- Rubinov, M. & Sporns, O. Complex network measures of brain connectivity: uses and interpretations. Neuroimage 52, 1059–1069 (2010).
- Minamisawa, G., Kwon, S. E., Chevée, M., Brown, S. P. & O'Connor, D. H. A Non-canonical feedback circuit for rapid interactions between somatosensory cortices. *Cell Rep.* 23, 2718–2731.e6 (2018).
- Li, A. et al. Micro-optical sectioning tomography to obtain a high-resolution atlas of the mouse brain. Science 330, 1404–1408 (2010).
- Wang, Y., Ye, M., Kuang, X., Li, Y. & Hu, S. A simplified morphological classification scheme for pyramidal cells in six layers of primary somatosensory cortex of juvenile rats. IBRO Rep. 5, 74–90 (2018).
- Phillips, J. W. et al. A repeated molecular architecture across thalamic pathways. Nat. Neurosci. https://doi.org/10.1038/s41593-019-0483-3 (2019).
- Huh, C. Y. L., Peach, J. P., Bennett, C., Vega, R. M. & Hestrin, S. Feature-specific organization of feedback pathways in mouse visual cortex. *Curr. Biol.* 28, 114–120.e5 (2018).
- Coogan, T. A. & Burkhalter, A. Hierarchical organization of areas in rat visual cortex. J. Neurosci. 13, 3749–3772 (1993).
- Crick, F. & Koch, C. Constraints on cortical and thalamic projections: the no-strong-loops hypothesis. Nature 391, 245–250 (1998).
- D'Souza, R. D., Meier, A. M., Bista, P., Wang, Q. & Burkhalter, A. Recruitment of inhibition and excitation across mouse visual cortex depends on the hierarchy of interconnecting areas. eLife 5, e19332 (2016).
- Bohland, J. W. et al. A proposal for a coordinated effort for the determination of brainwide neuroanatomical connectivity in model organisms at a mesoscopic scale. PLOS Comput. Biol. 5. e1000334 (2009).
- Han, Y. et al. The logic of single-cell projections from visual cortex. *Nature* 556, 51–56 (2018).
- Economo, M. N. et al. A platform for brain-wide imaging and reconstruction of individual neurons. eLife 5, e10566 (2016).
- Winnubst, J. et al. Reconstruction of 1,000 projection neurons reveals new cell types and organization of long-range connectivity in the mouse brain. Cell 179, 268–281.e13 (2019)
- Halassa, M. M. & Sherman, S. M. Thalamocortical circuit motifs: a general framework. Neuron 103, 762–770 (2019).
- Douglas, R. J. & Martin, K. A. C. Neuronal circuits of the neocortex. *Annu. Rev. Neurosci.* 27, 419–451 (2004).
- Gămănuţ, R. et al. The mouse cortical connectome, characterized by an ultra-dense cortical graph, maintains specificity by distinct connectivity profiles. Neuron 97, 698–715. e10 (2018).
- Hegdé, J. & Felleman, D. J. Reappraising the functional implications of the primate visual anatomical hierarchy. Neuroscientist 13, 416–421 (2007).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

#### Methods

#### Mice

Experiments involving mice were approved by the Institutional Animal Care and Use Committees of the Allen Institute for Brain Science in accordance with NIH guidelines. Sources of mouse lines are listed in Supplementary Table 1. Transgene expression patterns in many Cre driver lines used in this study have been previously characterized and are available through the Transgenic Characterization data portal (http://connectivity.brain-map.org/transgenic). Cre lines were originally derived on various backgrounds, but the majority were crossed to C57BL/6J mice for more than ten generations and maintained as heterozygous lines upon arrival. Tracer injections were performed in male and female mice at an average age of P56 + 10 days. Mice were grouphoused with a 12-h light–dark cycle. Food and water were provided ad libitum. No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

#### **Tracers and injection methods**

rAAV was used as an anterograde tracer. For most regions, stereotaxic coordinates were used to identify the appropriate location for a tracer injection. Atlas-derived stereotaxic coordinates were chosen for each target area based on The Mouse Brain in Stereotaxic Coordinates<sup>47</sup>. Anterior-posterior (AP) coordinates are referenced from bregma, medial-lateral (ML) coordinates are distance from midline at bregma, and dorsal-ventral (DV) depth is measured from the pial surface of the brain. Stereotaxic coordinates used for each experiment can be found through the data portal. For a subset of experiments in the left hemisphere, we first functionally mapped the visual cortex using intrinsic signal imaging (ISI) through the skull, described below to assist in targeting injections. A pan-neuronal AAV expressing EGFP (rAAV2/1. hSynapsin.EGFP.WPRE.bGH, Penn Vector Core, AV-1-PV1696, Addgene ID 105539) was used for injections into wild-type C57BL/6J mice (stock no. 00064, The Jackson Laboratory). To label genetically defined populations of neurons, we used either a Cre-dependent AAV vector that robustly expresses EGFP within the cytoplasm of Cre-expressing infected neurons (AAV2/1.pCAG.FLEX.EGFP.WPRE.bGH, Penn Vector Core, AV-1-ALL854, Addgene ID 51502) or, a Cre-dependent AAV virus expressing a synaptophysin-EGFP fusion protein to more specifically label presynaptic terminals (AAV2/1.pCAG.FLEX.sypEGFP.WPRE.bGH, Penn Vector Core).

Functional mapping of visual field space by ISI was used in some cases to guide injection placement. Additional details of this procedure can be found at http://help.brain-map.org/display/mouseconnectivity/Documentation?preview=/2818171/10813533/Connectivity\_Overview.pdf). In brief, a custom 3D-printed headframe was attached to the skull, centred at 3.1 mm lateral and 1.3 mm anterior to lambda on the left hemisphere. A transcranial window was made by securing a 7-mm glass coverslip onto the skull in the centre of the headframe well. Mice were allowed to recover for at least seven days before ISI mapping. ISI was then used to measure the haemodynamic response to visual stimulation across the entire field of view of a lightly anaesthetized, head-fixed, mouse. The visual stimulus consisted of sweeping a bar containing a flickering black-and-white checkerboard pattern across a grey background<sup>48</sup>. To generate a map, the bar was swept across the screen ten times in each of the four cardinal directions, moving at 9° per second. Processing of sign maps followed methods previously described<sup>49</sup>, with minor modifications. Phase maps were generated by calculating the phase angle of the pre-processed discrete Fourier transform at the stimulus frequency. The phase maps were used to translate the location of a visual stimulus displayed on the retina to a spatial location on the cortex. A sign map was produced from the phase maps by taking the sign of the angle between the altitude and azimuth map gradients. Averaged sign maps were produced from a

minimum of three time series images, for a combined minimum average of 30 stimulus sweeps in each direction. Visual area segmentation and identification was obtained by converting the visual field map into a binary image using a manually defined threshold and further processing the initial visual areas with a split/merge routine  $^{49}$ . Sign maps were curated and the experiment repeated if (1) fewer than six visual areas were positively identified; (2) retinotopic metrics of VISp were out of bounds (azimuth coverage within  $60-100^{\circ}$  and altitude coverage within  $35-60^{\circ}$ ); or (3) auto-segmented maps needed to be annotated with more than three adjustments. Each animal had three attempts to get a passing map.

ISI images were acquired using a pair of Nikon lenses (Nikkor 135 mm f/2.8 lens and 50 mm f/1.8), providing a magnification of 2.7×. Illumination was from a ring of sequential and independent LED lights, with green (peak wavelength of 527 nm and full width half maximum (FWHM) of 50 nm; Cree, C503B-GCN-CY0CO791) and red spectra (peak wavelength of 635 nm and FWHM of 20 nm; Avago Technologies, HLMP-EG08-Y2000), via a bandpass filter (630/92 nm, Semrock, FF01), and acquired with a sCMOS camera (Andor, Zyla 5.510-tap). Illumination and image acquisition were controlled with in-house GUI software written in Python. An image of the surface vasculature was acquired with green LED illumination to provide fiduciary marker references on the surface of the brain.

All mice received one unilateral injection into a single target region. For injections using stereotaxic coordinates from bregma as a registration point, procedures were followed as previously described¹. For ISI-guided injections, the glass coverslip of the transcranial window was removed by drilling around the edges and a small burr hole drilled, first through the Metabond and then through the skull using surface vasculature fiducials obtained from the ISI session as a guide. An overlay of the sign map over the vasculature fiducials was used to identify the target injection site. rAAV was delivered by iontophoresis with current settings of 3  $\mu A$  at 7 s 'on' and 7 s 'off' cycles for 5 min total, using glass pipettes (inner tip diameters of 10–20  $\mu m$ ).

Some injections were done into lines with regulatable versions of Cre. Tamoxifen-inducible Cre line (CreER) mice were treated with 0.2 mg/g body weight of tamoxifen solution in corn oil via oral gavage once per day for 5 consecutive days starting the week after virus injection. Trimethoprim-inducible Cre line (dCre) mice were treated with 0.3 mg/g body weight of trimethoprim solution in 10% DMSO via oral gavage once per day for 3 consecutive days starting the week after virus injection. For these Cre lines, brains were removed 4 weeks after the rAAV injection date as opposed to 3 weeks. All mice were deeply anaesthetized before intracardial perfusion, brain dissection, and tissue preparation for serial imaging as previously described¹.

## Serial two-photon tomography and image data processing

Imaging by STPT (TissueCyte 1000, TissueVision Inc. Somerville, MA) has been described  $^{1.50}$ , and here we used the exact same procedures as in our earlier published studies  $^{1.51}$ . In brief, following tracer injections, brains were imaged using STPT at high x-y resolution (0.35  $\mu m \times 0.35 \, \mu m$ ) every 100  $\mu m$  along the rostrocaudal z-axis, after which the images underwent quality control and manual annotation of injection sites, followed by signal detection and registration to the Allen Mouse Brain Common Coordinate Framework, version 3 (CCFv3) through our informatics data pipeline (IDP).

The IDP manages the processing and organization of the image and quantified data for analysis and display in the web application, as previously described  $^{1,52}$ . The two key algorithms are signal detection and image registration. Previous methods were implemented, except that two variations of the segmentation algorithm were employed, depending on the virus used for that experiment; one was tuned for EGFP detection and one for SypEGFP detection. High-threshold edge information was combined with spatial distance-conditioned low-threshold edge results to form candidate signal object sets. The candidate objects

were then filtered on the basis of morphological attributes such as length and area using connected component labelling. For the SypEGFP data, filters were tuned to detect smaller objects (punctate terminal boutons versus long fibres). In addition, high-intensity pixels near the detected objects were included in the signal pixel set. Detected objects near hyper-intense artefacts that occurred in multiple channels were removed. We developed an additional filtering step using a supervised decision tree classifier to filter out surface segmentation artefacts, based on morphological measurements, location context and the normalized intensities of all three channels.

The output is a full resolution mask that classifies each 0.35  $\mu m \times 0.35 \, \mu m$  pixel as either signal or background. An isotropic 3D summary of each brain is constructed by dividing each image series into 10  $\mu m \times 10 \, \mu m \times 10 \, \mu m$  grid voxels. Total signal is computed for each voxel by summing the number of signal-positive pixels in that voxel. Each image stack is registered in a multi-step process using both global affine and local deformable registration to the 3D Allen mouse brain reference atlas as previously described 52. Segmentation and registration results are combined to quantify signal for each voxel in the reference space and for each structure in the reference atlas ontology by combining voxels from the same structure.

Once an image series had passed quality control steps, injection site polygons overlaying the cell bodies of infected neurons were manually drawn. These polygons were informatically warped into the CCFv3 atlas space. Green channel signal intensity within the polygons was used to identify which structures have been injected, and to quantify the relative magnitude of their infections. The structure that received the largest proportion of signal intensity was identified as the primary injection site structure, and all other structures were considered secondary structures containing infected cells. A quantified injection summary is provided for each image series through the data portal that shows the relative amount of signal detected within each infected structure.

# $\label{thm:projection} Quantification of projection strengths using segmentation and registration$

Projection signals can be quantified in several ways using our informatics pipeline (see SDK help: https://allensdk.readthedocs.io/en/latest/connectivity.html#structure-level-projection-data). Here, we most frequently report 'normalized projection volume', which is the volume of detected projection signals in all voxels in a structure (in mm³), divided by the total volume of detected signal in the manually annotated injection site. We also use the 'normalized connection densities' output from the voxel-level interpolation model for modularity analyses in Fig. 1e. Connection density is the sum of detected projection pixels divided by the sum of all pixels in that voxel or structure. Normalized connection density is this value divided by the injection site density.

It is important to note that even after undergoing our quality control procedures, these informatically derived measures of connection strength can include artefacts (false positives), and, particularly for the EGFP tracer, report total signal from labelled axons, including passing fibres and synaptic terminals. For this reason, we performed extensive manual checking of all CC, CT, and TC targets to remove any signals from regions in which we could not identify any true positive axons or terminals (see main text).

## Morphological reconstruction of single L4 neurons

The Cux2-IRES-CreERT2 driver line was crossed with a TIGRE2.0 reporter line<sup>27</sup>, Ai166, also known as TIGRE-MORF<sup>53</sup>. In brief, Ai166 expresses a Cre-dependent MORF transgene, composed of a farnesylated EGFP preceded by a stretch of 22 guanidine nucleotides (22G-GFPf), which puts the transgene out of frame. Rare DNA replication errors lead to the deletion of one G, correcting the frameshift, and leading to GFPf expression. Combining Ai166 with a CreERT2 line and giving mice a low dose of tamoxifen produces sparse cellular

labelling that is well suited for 3D morphological reconstruction  $^{53}$ . High-resolution whole-brain imaging by fMOST has been described previously  $^{54}$ ; similar protocols were used here to image the Cux2-IRES-CreERT2;Ai166 brain. Specifically, high-resolution block-face fluorescence imaging was done in coronal planes. Using a diamond knife, 1.0-µm sections were removed before we imaged subsequent planes. The process was repeated through the entire rostral–caudal extent of the mouse brain, producing more than 10,000 images with a resolution of  $0.3\times0.3\times1\,\mu\text{m}$  ( $x\times y\times z$ ). Following acquisition of the complete fMOST image stack, it was converted into a multi-level navigable data set using the Vaa3D-TeraFly program  $^{55}$ , and then reconstructions were performed using Vaa3D-TeraVR software tools built to facilitate semi-automated and manual reconstructions  $^{56}$ .

# Creation of the cortical top-down and flattened views of the CCFv3 for data visualization

A standard z-projection of signal in a top-down view of the cortex mixes signal from multiple areas. Visualizations of fluorescence in Figs. 1-3 instead project signal along a curved cortical coordinate system that more closely matches the columnar structure of the cortex. This coordinate system was created by first solving Laplace's equation between pia and white matter surfaces, resulting in intermediate equi-potential surfaces. Streamlines were computed by finding orthogonal (steepest descent) paths through the equi-potential field. Cortical signal can then be projected along these streamlines for visualization.

A cortical flatmap was also constructed to enable visualization of anatomical and projection information while preserving spatial context for the entire cortex. The flatmap was created by computing the geodesic distance (the shortest path between two points on a curve surface) between every point on the cortical surface and two pairs of selected anchor points. Each pair of anchor points forms one axis of the 2D embedding of the cortex into a flatmap. The 2D coordinate for each point on the cortical surface is obtained by finding the location such that the radial (circular) distance from the anchor points (in 2D) equals the geodesic distance that was computed in 3D. This procedure produces smooth mapping of the cortical surface onto a 2D plane for visualization. This embedding does not preserve area and the frontal pole and medial-posterior region is highly distorted. As such, all numerical computation is done in 3D space. Similar techniques are used for texture mapping on geometric models in the field of computer graphics<sup>57</sup>.

#### Network modularity analysis

The matrix of connection weights between cortical areas (Fig. 1e) was obtained from a model of voxel-level connectivity  $^{29}$ . We analysed the network structure of this graph using the Louvain Community Detection algorithm from the Brain Connectivity Toolbox (https://sites.google.com/site/bctnet/) $^{30,58}$ . We determined the modularity metric (Q) at various levels of granularity by varying the resolution parameter,  $\gamma$ , from 0 to 2.5 in steps of 0.1. Q quantifies the fraction of connections inside modules minus the fraction of connections expected inside the same modules if the network was connected randomly; that is, Q=0 has no more intramodule connections than expected by chance, while Q>0 indicates a network with some community structure.

For each value of  $\gamma$ , the modularity was computed 1,000× and each pair of regions received an affinity score between 0 and 1. The affinity score is the probability of two regions being assigned to the same module weighted by the modularity score (Q) for that iteration, thereby assigning higher weights to partitions with a higher modularity score. Each region was assigned to the module with which it had the highest affinity, with the caveat that all structures within a module had an affinity score  $\geq 0.5$  with all other members of the module. For each value of  $\gamma$ , we also generated a shuffled matrix containing the same weights but with the source and target regions randomized. The modularity for the cortical matrix (Q) and the shuffled matrix (Q<sub>shuffled</sub>) were evaluated at

each value of  $\gamma$ . As stated in the results, we chose to focus on the modules identified at  $\gamma=1.3$  (Q=0.36) where the difference between Q and  $Q_{\text{shuffled}}$  was at its peak ( $0.22\pm0.017$ , mean  $\pm$  s.d.), although it should be noted that it was relatively stable between  $\gamma=1$  and  $\gamma=1.8$  ( $0.21\pm0.019$  at  $\gamma=1$ ,  $0.20\pm0.012$  at  $\gamma=1.8$ ). Modules were identical from  $\gamma=1.3$  to  $\gamma=1.5$  and showed only minor differences for  $\gamma$  between 1 and 2.

#### Statistics and reproducibility

We used the software program GraphPad Prism for statistical tests and generation of graphs, and the software program Gephi for visualization and layout of network diagrams<sup>59,60</sup>. The exact numbers of tracer injection experiments per mouse line and source area are shown in Supplementary Table 1, and range from n = 1 to 31. Not all experiments were independently repeated because we sought to balance the need for broad coverage across Cre lines and source areas with excessive animal use. Previously, we demonstrated that n=1 is a good predictor of connectivity strengths across multiple animals<sup>1</sup>. In this study, we also show that the correlations between brain-wide projection strengths from experiments at matched locations within the same mouse line are consistent, positive, and significant (Spearman r > 0.8, P < 0.0001, Extended Data Fig. 1). Sample sizes for analyses presented in all figures are mostly noted in the main text, and can also be found in associated Supplementary Tables. Specifics include: for Fig. 2g, h: n = number of mice per line, for wild-type, Cux2, Rbp4: n = 27, Syt6: n = 23, A93: n = 22, Tlx3, Ntsr1: n = 21, Scnn1a-Tg3: n = 19, Chrna2, Efr3a: n = 15, Nr5a1: n = 10, Sim1: n = 9, Rorb: n = 6, Sepw1: n = 5; for Fig. 4c: n = number of mice per line, for wild-type, Rbp4: n = 27, Syt6: n = 23, and Ntsr1: n = 21; for Fig. 4d: n = 1,158 total CT connections, 462 are shared above threshold, for Fig. 4e: n = 1.892 total CT connections. 628 are shared above threshold. for Fig. 4f: n = 1,158 total CT connections, 495 are shared above threshold; for Fig. 5a: n = 7,063 unique connections (columns). Numbers of replicate experiments per each of the 7,063 connections ranged from 1 to 53, and are listed in Supplementary Table 8; for Fig. 5f: the number of connections assigned to each cluster is plotted in Fig. 5c, and can also be found in Supplementary Table 8 (cluster 1: n = 1,740, cluster 2: n = 366, cluster 3: n = 375, cluster 4: n = 228, cluster 5: n = 602, cluster 6: n = 2,224, cluster 7: n = 102, cluster 8: n = 129, cluster 9: n = 1,297). The number of connections per cortical Cre line can also be found in Supplementary Table 8, for A93: n = 375, C57Bl6/J/Emx1: n = 1,431, Chrna2: n = 136, Cux2: n = 703, Efr3a: n = 223, Nr5a1: n = 251, Ntsr1: n = 246, Rbp4: n = 1,149, Rorb: n = 185: Scnn1a-Tg3: n = 263. Sepw1: n = 140. Sim1: n = 108. Svt6: n = 150. Tlx3: n=1.043), and per thalamic projection class was, for core: n=62. matrix-focal: n = 136, intralaminar: n = 160, matrix-multiareal: n = 302. Figure 6b: n = 385 total CT connections.

#### Clustering analyses

Unsupervised hierarchical clustering was conducted with the online software, Morpheus (https://software.broadinstitute.org/morpheus/). log-transforms were calculated on all values after adding a small value (0.5 minimum of the true positive array elements) to avoid log (0). Proximity between clusters was computed using average linkages with Spearman rank correlations as the distance metric. Relative layer density is the fraction of the total projection signal in each layer, scaled by the relative layer volumes in that target. The clustering algorithm works agglomeratively: initially assigning each sample to its own cluster and iteratively merging the most proximal pair of clusters until finally all the clusters have been merged. To compare distances between granular and agranular samples (those that lack L4), we used the median of the other present layers for L4.

#### Unsupervised discovery of hierarchy position

Following the classification of the laminar patterns into nine clusters of CC and TC connections, we used an unsupervised method to simultaneously assign a direction to a cluster type and to construct a hierarchy.

We first defined hierarchy scores of cortical regions based on layer-termination patterns of CC connections. First consider a mapping function  $M_{CC}$  for CC connections:

$$M_{\rm CC}$$
:  $\{1, ..., 9\} \rightarrow \{-1, 1\}$  (1)

which maps a type of connection cluster ( $C_{T_{i,j}} \in \{1,...,9\}$ , where  $C_{T_{i,j}}$  denotes the layer termination pattern of the connection from area j to area i for Cre line T) to either feedforward ( $M_{\rm CC}=1$ ) or feedback ( $M_{\rm CC}=-1$ ) type. We search over the space of possible maps to see which map produces the most self-consistent hierarchy. As some transgenic lines have different numbers of connections in different clusters, some maps will lead to particular transgenic lines having very biased feedforward or feedback calls. Thus, we add a confidence measure (conf(T)) for each Cre line T, which decreases the importance of the information provided by a transgenic line to the CC global hierarchy if the calls from that transgenic line are biased. This allows us to reduce the bias in the regions where experiments used more Cre lines that predominantly mark feedforward or feedback connections. The Cre-dependent confidence measure is defined as:

$$\operatorname{conf}(T) = 1 - |\langle M_{CC}(C_{T_{i,j}}) \rangle_{i,j}| \tag{2}$$

with a global confidence as an average over all the inter-areal connections above the threshold ( $10^{-1.5}$ )

$$confg = \langle conf(T) \rangle_{i,i}$$
 (3)

We define the initial hierarchical position of an area as:

$$H_i^0 = \frac{1}{2} \left( \left\langle M_{CC} \left( C_{T_{i,j}} \right) \times \text{conf}(T) \right\rangle_j - \left\langle M_{CC} \left( C_{T_{j,i}} \right) \times \text{conf}(T) \right\rangle_j \right)$$
(4)

The first term,  $\langle M_{CC}(C_{T_{i,j}}) \times \operatorname{conf}(T) \rangle_j$ , describes the average direction of connections to area i, and thus represents the hierarchical position of the area as a target. The second term,  $-\langle M_{CC}(C_{T_{j,i}}) \times \operatorname{conf}(T) \rangle_j$ , on the other hand, represents the average direction of connections from area i, depicting the hierarchical position of the area as a source. The hierarchical position of a cortical area is the average between its hierarchical position as source and target.

To test how self-consistent a hierarchy is we define the CC global hierarchy score:

$$h_{\text{CC}} = \frac{1}{\text{confg}^2} \langle M_{\text{CC}}(C_{T_{i,j}}) \times \text{conf}(T) \times (H_i^0 - H_j^0) \rangle_{i,j}$$
 (5)

We performed an exhaustive search over all the maps  $M_{\rm CC}$  for the entire set of CC connections, and the most self-consistent hierarchy that maximizes the CC global hierarchy score is obtained when connections of clusters 2, 6 and 9 are of one direction and 1, 3, 4, 5, 7 and 8 are of the opposite direction. As described in the main text, we conclude that clusters 2, 6 and 9 are feedback connection patterns, and the other group of clusters corresponds to feedforward.

The initial hierarchy score  $(H_i^0)$  of each area i is thus obtained by computing the average direction of connections to and from the area (Eq. (4)) while concurrently searching for the optimal mapping of each lamination pattern to either the feedforward or feedback direction, and is bounded by -1 and 1. After we obtain the initial positions in the hierarchy, the hierarchy scores of all cortical regions are iterated until the fixed points are reached, to refine the cortical hierarchy. Without iterations, the hierarchy scores account only for the number of feedforward and feedback connections each area receives or sends out. Therefore, the initial hierarchy obtained by Eq. (4) alone does not account for the hierarchy positions of the target and source areas that each cortical area makes connections to, and places any two areas with

the same number of feedforward and feedback connections at the same level in the hierarchy. To address this issue, we implement a two-step iterative scheme:

$$H_i^{n-1/2} = \frac{1}{2} \{ \langle H_j^{n-1} + M_{CC}(C_{T_{i,j}}) \rangle_j - \langle -H_j^{n-1} + M_{CC}(C_{T_{j,i}}) \rangle_j \}$$
 (6)

$$H_i^n = H_i^{n - \frac{1}{2}} - \left\langle H_j^{n - \frac{1}{2}} \right\rangle_i \tag{7}$$

where n refers to iterative steps. The first part (Eq. (6)) refines the hierarchy score of area i on the basis of the current hierarchy scores of its target and source areas. The next part (Eq. (7)) subtracts the hierarchy scores averaged over all areas to remove possible drifts. At every iteration step we also check to see whether the mapping of connection clusters to feedforward or feedback connection needs to change; however, it remained constant through the iterations. We found that the hierarchy scores reach the fixed points after just a few (<5) iterations, and used 20 iterations to find the final hierarchy scores of all areas. These final hierarchy scores are denoted as the hierarchy obtained by CC connections.

Next, we examined whether and how the TC connections affect the cortical hierarchy, by incorporating layer termination patterns of TC connections in addition to the CC connections. As in CC connections, TC connections are clustered into nine types on the basis of their layer termination patterns. The mapping of the lamination patterns is based on the hierarchy scores of cortical regions obtained by CC connections, while the hierarchical positions of thalamic areas relative to the cortical areas are found concurrently. The mapping of the TC layer termination types to directions is defined as:

$$M_{TC}: \{1, ..., 9\} \to \{-1, 1\}$$
 (8)

similar to the mapping of CC connections in Eq. (1). Because thalamic areas are always the source in TC connections, the initial hierarchy score of each thalamic area i is defined by the average direction of connections from the area:

$$H_i^0 = -\left\langle M_{\text{TC}}(C_{T_{j,i}}) \times \frac{\min(N_{\text{ff}}, N_{\text{fb}})}{N_{\text{ff}} + N_{\text{fb}}} \right\rangle_j \tag{9}$$

where the mapping of the lamination patterns,  $M_{TC}$  is obtained by searching for the most self-consistent assignment that maximizes the TC global hierarchy score  $h_{TC}$ :

$$h_{\text{TC}} = \left\langle M_{\text{TC}}(C_{T_{i,j}}) \times (H_i^0 - H_j^0) \times \frac{\min(N_{\text{ff}}, N_{\text{fb}})}{N_{\text{ff}} + N_{\text{fb}}} \right\rangle_{i,i}$$
(10)

The parameters  $N_{\rm ff}$  and  $N_{\rm fb}$  refer to the numbers of feedforward and feedback TC connections, respectively. The multiplier  $\frac{\min(N_{\rm ff},N_{\rm fb})}{N_{\rm ff}+N_{\rm fb}}$  biases the optimization method to preferentially search for mappings that result in roughly equal numbers of feedforward and feedback connections. Without such a weight on equal divide of the connections, the search algorithm decides TC connections to be always feedforward, placing all thalamic areas below cortical areas.

As with CC connections, we performed an exhaustive search over all the maps  $M_{\rm TC}$  for the entire set of TC connections to find the most self-consistent hierarchy that maximizes the TC global hierarchy score. For TC connections, we found that connections of cluster 2 and 6 are of one direction and the rest of the clusters are of the opposite direction. Again, as described in the main text, we conclude that clusters 2 and 6 are feedback, and the rest correspond to feedforward patterns.

Once the initial positions of the thalamic areas in the hierarchy are obtained, hierarchy scores of thalamic and cortical areas are iterated

until the fixed points are reached (20 iterations), using a full mapping function  $M_{\rm CC+TC}$  that combines  $M_{\rm CC}$  and  $M_{\rm TC}$  for CC and TC connections, respectively:

$$H_i^{n-1/2} = \frac{1}{2} \{ (H_j^{n-1} + M_{\text{CC+TC}}(C_{T_{i,j}}))_j - (-H_j^{n-1} + M_{\text{CC+TC}}(C_{T_{j,i}}))_j \}$$
 (11)

$$H_i^n = H_i^{n - \frac{1}{2}} - \left\langle H_j^{n - \frac{1}{2}} \right\rangle_j \tag{12}$$

Finally, the effect of CT connections on the hierarchy is considered. Either feedforward or feedback direction is assigned to CT connections depending on the cortical layer from which the connections originate. Specifically, we classified CT connections based on the  $\log_{10}$ -transformed NPV from layers 5 and 6 of the source areas. Therefore, the mapping of CT connections is described by:

$$M_{\text{CT}}$$
: [L5 log<sub>10</sub>NPV, L6 log<sub>10</sub>NPV]  $\rightarrow \{-1, 1\}$  (13)

We first determined the predicted direction (feedforward or feedback) of each CT connection based on the hierarchy constructed from CC and TC projection patterns. These directions of CT connections show mixed L5 and L6 source expressions. To classify the CT connections to either L5 or L6 dominance and, subsequently, to feedforward or feedback, we used linear discriminant analysis on  $\log_{10}$ -transformed NPV values of L5 and L6 lines with a prior that biases the method to yield about equal numbers of L5 and L6-dominant connections. The classifier assigns feedforward direction to connections with stronger L5 source NPV, and feedback direction to L6 dominant connections, using the linear separator. Once directions of CT connections have been obtained, the mappings  $M_{\rm CC}$ ,  $M_{\rm TC}$  and  $M_{\rm CT}$  are combined to construct a comprehensive mapping  $M_{\rm CC+TC+CT}$  of all connections among cortical and thalamic areas to directions. The initial positions of thalamic regions in the hierarchy are computed by:

$$H_i^0 = \frac{1}{2} (\langle M_{\text{TC+CT}}(C_{T_{i,j}}) \rangle_j - \langle M_{\text{TC+CT}}(C_{T_{j,i}}) \rangle_j)$$
 (14)

where  $M_{\mathsf{TC+CT}}$  is the mapping of all TC and CT connections. Note that the multiplier  $\frac{\min(N_{\mathsf{ff}},N_{\mathsf{fb}})}{N_{\mathsf{ff}}+N_{\mathsf{fb}}}$  used for initial thalamic hierarchy with TC connections only (which biases thalamus to be towards the centre of the hierarchy) is not needed here, owing to the presence of the CT connections in the computations. However, the bias is not fully eliminated as it influenced the initial assignment of CT and TC connections types to be feedforward or feedback. The initial hierarchy scores are iterated together with hierarchy scores of cortical areas obtained from Eqs. (6, 7):

$$H_i^{n-1/2} = \frac{1}{2} \{ \langle H_j^{n-1} + M_{\text{CC+TC+CT}}(C_{T_{i,j}}) \rangle_j - \langle -H_j^{n-1} + M_{\text{CC+TC+CT}}(C_{T_{j,i}}) \rangle_j \}$$
 (15)

$$H_{i}^{n} = H_{i}^{n - \frac{1}{2}} - \left\langle H_{j}^{n - \frac{1}{2}} \right\rangle_{i} \tag{16}$$

In this way, we obtained three different versions of cortical hierarchy constructed from: (1) CC connections only, (2) CC connections and thalamocortical connections, and (3) CC, TC, and CT connections.

We examined how the additional information provided by TC and CT connections affects the self-consistency of the hierarchy by comparing the global hierarchy scores of the three different versions of hierarchy. For this purpose, we compared the global hierarchy scores without any confidence or weight multiplier:

$$h = \left\langle M\left(C_{T_{i,j}}\right) \times \left(H_i - H_j\right)\right\rangle_{i,j} \tag{17}$$

In addition to the hierarchy of all areas, we also constructed the intermodule hierarchy of cortical areas. We used the same mappings obtained from construction of the all-area hierarchy to classify the lamination patterns. For intermodule hierarchy, all the connections to and from each module were used to build the hierarchy among the modules.

# Global hierarchy score of shuffled connectomes and 'perfectly hierarchical' connectome

To evaluate 'how hierarchical' the mouse brain is, we generated shuffled connectivity data for the connection patterns, computed the global hierarchy scores, and compared the global hierarchy scores of the shuffled connectomes to that of the mouse brain connectome. The shuffled connectivity is constructed by randomly rearranging sources and targets, while preserving the projection layer patterns and the distributions of source and target areas, within each Cre line. We generated 100 versions of shuffled connectivity data, and calculated their global hierarchy scores as was done with the original connectivity data, described in the previous section. The medians of the shuffled distributions provide an estimate of the lower bound of this score (0.001, 0.044, -0.002, for CC, CC+TC, CC+TC+CT, respectively; Fig. 6c).

We also generated connectivity data with perfectly self-consistent hierarchy, which provides the upper bound of the global hierarchy score. To do this, we assigned a direction (feedforward or feedback) for each connection in the mouse brain connectivity data, based on the final hierarchy positions of the cortical and thalamic regions. With this 'true' mapping of each connection to a direction, the global hierarchy score is computed using Eq. (17), producing values of 0.679, 0.636, and 0.683, respectively, for CC, CC + TC, and CC + TC + CT connections.

Therefore, comparison of global hierarchy scores allows us to evaluate how hierarchical the mouse brain is compared to the hierarchy by chance (shuffled) and the perfect hierarchy (upper bound). The global hierarchy scores with the shuffled mean subtracted and normalized by the strictly hierarchical data provides a single measure that quantifies the steepness of hierarchy across arbitrarily different connectivity data.

#### **Reporting summary**

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

#### **Data availability**

Data (including high-resolution images, segmentation, registration to CCFv3, and automated quantification of injection size, location, and distribution across brain structures) are available through the Allen Mouse Brain Connectivity Atlas portal (http://connectivity.brain-map.org/). Individual experiment summaries can be viewed using this link: http://connectivity.brain-map.org/projection/experiment/[insert experimental id]. Experimental ids are listed in Supplementary Table 2. In addition to visualization and search tools available at this site, users can download data using the Allen Brain Atlas API (http://help.brain-map.org/display/mouseconnectivity/API) and the Allen Brain Atlas Software Development Kit (SDK: http://alleninstitute.github.io/AllenSDK/connectivity.html). Through the SDK, structure and voxel-level projection data are available for download. Examples of code for common

data requests are provided as part of the Mouse Connectivity Jupyter notebook to help users get started with their own analyses. Source data generated for this study are provided as Supplementary Tables as indicated throughout. Code and data files for hierarchical analyses are available through the Allen SDK and Github (https://github.com/AllenInstitute/MouseBrainHierarchy).

- Franklin, K. B. J. & Paxinos, G. The Mouse Brain in Stereotaxic Coordinates (Elsevier Academic, 2012).
- Kalatsky, V. A. & Stryker, M. P. New paradigm for optical imaging: temporally encoded maps of intrinsic signal. Neuron 38, 529–545 (2003).
- Garrett, M. E., Nauhaus, I., Marshel, J. H. & Callaway, E. M. Topography and areal organization of mouse visual cortex. J. Neurosci. 34, 12587–12600 (2014).
- Ragan, T. et al. Serial two-photon tomography for automated ex vivo mouse brain imaging. Nat. Methods 9, 255–258 (2012).
- Martersteck, E. M. et al. Diverse central projection patterns of retinal ganglion cells. Cell Rep. 18, 2058–2072 (2017).
- Kuan, L. et al. Neuroinformatics of the Allen mouse brain connectivity atlas. Methods 73, 4–17 (2015).
- 53. Wang, Y. et al. Complete single neuron reconstruction reveals morphological diversity in molecularly defined claustral and cortical neuron types. Preprint at https://www.biorxiv. org/content/10.1101/675280v1 (2019).lf ref. 53 (preprint) has now been published in final peer-reviewed form, please update the reference details if appropriate.
- Gong, H. et al. Continuously tracing brain-wide long-distance axonal projections in mice at a one-micron voxel resolution. *Neuroimage* 74, 87–98 (2013).
- Bria, A., Iannello, G., Onofri, L. & Peng, H. TeraFly: real-time three-dimensional visualization and annotation of terabytes of multidimensional volumetric images. *Nat. Methods* 13, 192-194 (2016).
- Wang, Y. et al. TeraVR empowers precise reconstruction of complete 3-D neuronal morphology in the whole brain. Nat. Commun. 10, 3474 (2019).
- Oliveira, G. N., Torchelsen, R. P., Comba, J. L. D., Walter, M. & Bastos, R. Geotextures: a multi-source geodesic distance field approach for procedural texturing of complex meshes. 2010 23rd SIBGRAPI Conf. Graphics, Patterns and Images 126–133 (IEEE, 2010).
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. 2008, P10008 (2008).
- Bastian, M., Heymann, S. & Jacomy, M. Gephi: an open source software for exploring and manipulating networks. Intl AAAI Conf. Weblogs and Social Media 3, 361–362 (2009).
- Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. PLoS One 9, e98679 (2014).

Acknowledgements We thank the Animal Care, Transgenic Colony Management and Laboratory Animal Services teams for mouse husbandry and tissue preparation; all the members of the Neurosurgery and Behavior team for viral injections, including those not listed as authors: N. Berbesque, N. Bowles, S. Cross, M. Edwards, S. Lambert, W. Liu, K. Mace, N. Mastan, C. Nayan, B. Rogers, J. Swapp, C. White and N. Wong; H. Gu for cloning of the synaptophysin–EGFP viral vector; E. Lee, F. Griffin and T. Nguyen for intrinsic signal imaging; and J. Royall and P. Lesnar for schematic figure preparation. This work was supported by the Allen Institute for Brain Science and, in part, by National Institutes of Health grants RO1AG047589 to J.A.H and U01MH105982 and U19MH114830 to H.Z. We thank the Allen Institute founder, Paul G. Allen, for his vision, encouragement, and support.

Author contributions Conceptualization: H.Z., J.A.H. and S. Mihalas. Supervision: H.Z., J.A.H., S. Mihalas, A.B., L.N., N. Gaudreault, P.A.G., J. Lecoq, S.A.S., J.W.P., A.R.J. and C.K. Project administration: S. McConoughey, S.W.O. and W.W. Investigation, validation, methodology and formal analyses: J.A.H., S. Mihalas, K.E.H., H.C., J.D.W., J.E.K., P.B., S.C., L.C., A.C., A.F., N. Gaudreault, N. Graddis, C.R.G., P.A.G., A.M.H., A.H., R.H., L.K., X.K., J. Lecoq, J. Luviano, P.L., Y.L., M.T.M., M.N., L.N., B.O., E.S., S.A.S., Q.W., A.W. and Y.W. Data curation: J.A.H., K.E.H., J.D.W., P.B., S.C., A.M.H., B.O. and W.W. Visualization: J.A.H., K.E.H., J.D.W., H.C., L.N., D.F., S. Mihalas, M.N. and Y.W. The original draft was written by J.A.H. with input from K.E.H., J.D.W., S. Mihalas, H.C., Q.W., C.K. and H.Z. All co-authors reviewed the manuscript.

Competing interests The authors declare no competing interests.

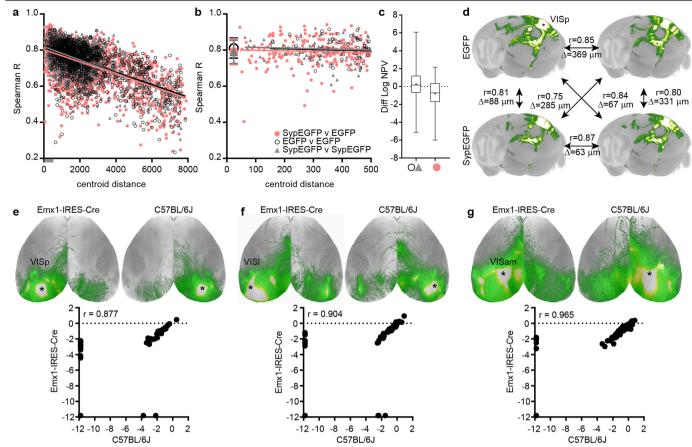
#### Additional information

**Supplementary information** is available for this paper at https://doi.org/10.1038/s41586-019-1716-z.

Correspondence and requests for materials should be addressed to J.A.H.

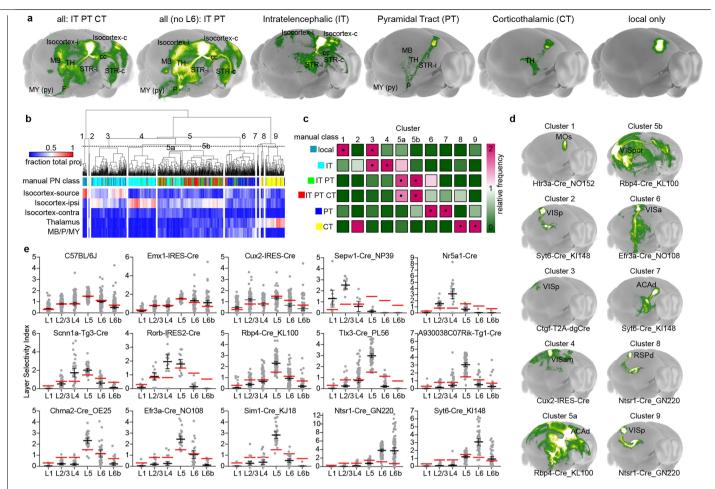
Peer review information *Nature* thanks Claus Hilgetag, Moritz Helmstaedter and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints.



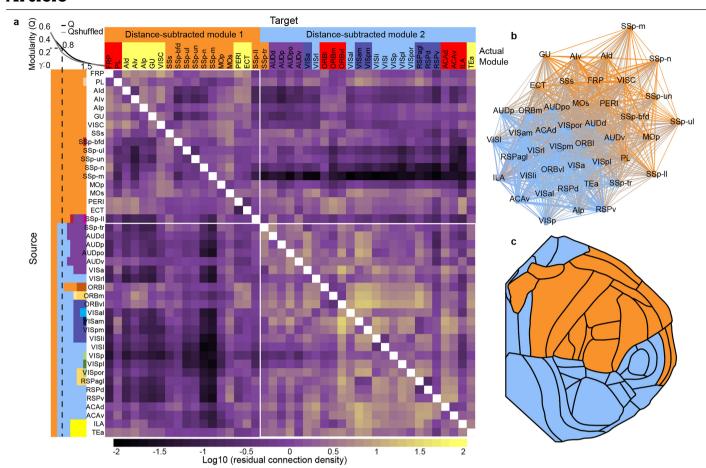
Extended Data Fig. 1 | Similarity of connection strengths by distance, virus, hemisphere, and Emx1-IRES-Cre or C57BL/6J mice. a-d, Most experiments were done with the Cre-dependent rAAV tracer, rAAV2/1.pCAG.FLEX.EGFP. WPRE. A subset of left hemisphere injections had a duplicate injection of rAAV with a synaptophysin-EGFP fusion transgene in place of the cytoplasmic EGFP (rAAV2/1.pCAG.FLEX.SypEGFP.WPRE). This tracer allowed us to address whether labelling presynaptic terminals would improve the accuracy with which we could quantify target connection strength, particularly in brain regions that contain mostly fibres of passage. Data consisted of n = 275experiments (137 EGFP, 138 SypEGFP). These were matched across Cre lines and areas, and represent n = 8 Cre lines and n = 26 cortical areas. For pairs of spatially matched experiments, the average projection strength (log<sub>10</sub>transformed NPV) measured across the entire brain was lower in SypEGFP than in EGFP experiments (~0.8 log unit when <500  $\mu$ m apart). However, brain-wide projection values were still highly and significantly correlated. Thus, we included the SypEGFP data sets when indicated for analyses of connectivity patterns from given source areas (but only in comparison with other SypEGFP data sets). a, Spearman correlation coefficients (r) of normalized projection volumes for all possible pairs of injections (different and same tracer, all in the same Creline) plotted against the distance between the injection centroids. Linear regressions showed a significant negative slope (P < 0.0001) with rdecreasing as distance between injections increased. **b**, *r* plotted for injections within 500 µm of each other; slopes were not significantly different from zero and means were not significantly different from each other. Average and s.d.

for each group is shown by the large symbols on the left (EGFP vs EGFP:  $0.81\pm0.056$ , SypEGFP vs SypEGFP:  $0.79\pm0.064$ , SypEGFP vs EGFP:  $0.79 \pm 0.071$ ). **c**, Quantitative differences in projection strengths measured between replicates with the same virus and between SypEGFP and EGFP (logNPV(EGFP) – logNPV(SypEGFP) injections, all < 500 μm apart in the same Cre line (n = 133 within virus and 222 between virus comparisons). Boxplots show median, IQR, minimum and maximum values; + indicates mean. d, Maximum intensity projections from four experiments within 500 µm of each other illustrate overall similarities between replicate injections and tracers (r shown for each pair). Injections targeted primary visual cortex (VISp) in Emx1-IRES-Cre mice using either EGFP or SypEGFP tracers as indicated. e-g, Injections into Emx1-IRES-Cre mice were made into visual areas on the left hemisphere, whereas all C57BL/6I mice received injections into the right hemisphere. Following registration to the CCF, which is a symmetric atlas, we identified three pairs of experiments in which the injection centroids were <500  $\mu$ m apart after we flipped injection site coordinates from the left to the right. Cortical projections were visually similar across both lines and hemispheres, and cortical connectivity strengths (to the 86 cortical targets) from these individual experiments (normalized projection volumes) were positively and strongly correlated as indicated. Thus, in Fig. 2 we merged the Emx1 and C57BL/6J data to represent connection strengths from all layers and classes, and in some of the 'anchor' groups we used data from both left and right hemisphere injections.



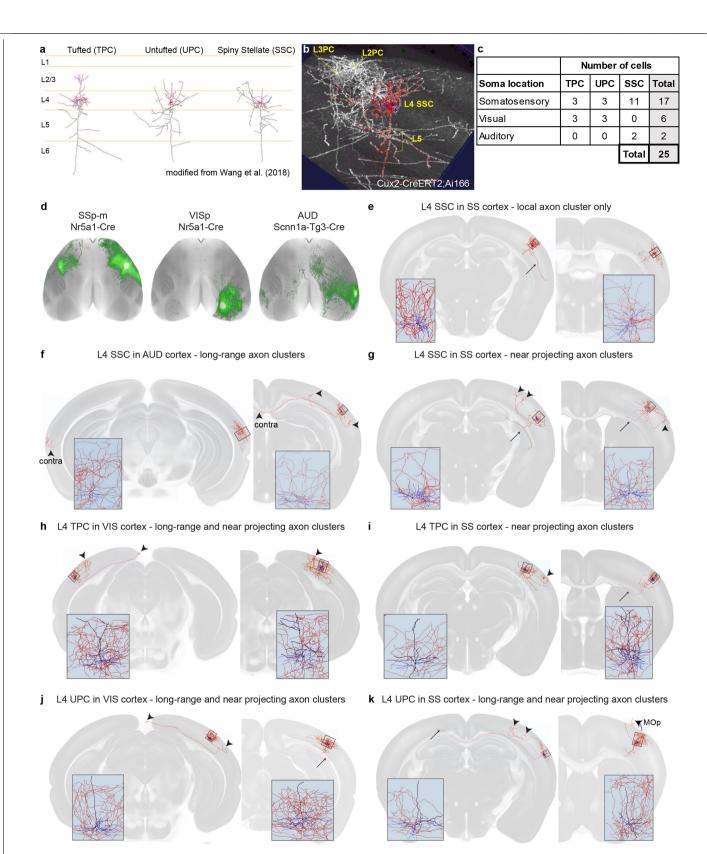
Extended Data Fig. 2 | Characterization of cortical projection neuron classes and layer selectivity across mouse lines. a, Brain-wide projection patterns were visually inspected for every experiment and manually classified into one of six categories on the basis of projections to ipsilateral and contralateral cortex, striatum, thalamus, and midbrain, pons or medulla structures as described for IT, PT, and CT classes. b-d, Unsupervised hierarchical clustering (using Euclidean distance and average linkage) of projection weights validates and reveals major classes of cortical projection neurons. **b**, Each column of the heat map shows one of the 1,081 injection experiments. Colours in the 'manual PN' class are coded as in **c** for projection class. Rows show selected major brain regions that distinguish known classes of projection neurons. Values in each cell are the fractions of total brain projection volume in the given region. The dendrogram was split into nine clusters, with two subclusters identified posthoc for cluster 5. The numbers of experiments per cluster were: 1, n = 24; 2, n = 4;3, n = 204; 4, n = 158; 5a, n = 148; 5b, n = 230; 6, n = 174; 7, n = 12; 8, n = 16; 9, n = 111.The numbers of experiments per projection class were: CT, n = 119; IT, n = 342; IT PT, n = 158; IT PT CT, n = 189; local, n = 100; PT, n = 173. **c**, The relative frequency of experiments from manually assigned projection classes within each cluster is shown. There was significant enrichment of 1, or 2 related, classes in each

cluster (dots; Fisher's exact t-test, P < 0.01). **d**, Maximum intensity projections of GFP-labelled axons across the brain from one example per cluster. e. Characterization of layer selectivity in wild-type mice and 14 Cre lines derived from injection experiments. Number of experiments per line is listed in Supplementary Table 1. For every injection and line, we assessed layer selectivity on the basis of the manually annotated injection sites. Polygons were drawn around every injection site so that, after registration to the CCF, injection volume in each layer could be informatically derived. A layerselectivity index was calculated for each experiment (the fraction of the total injection volume contained in each layer, scaled by the relative volume of each  $layer in the injection source \, region, because \, layer \, volumes \, differ \, by \, area). \, Plots \,$ show individual data points and the average layer selectivity index ± 95% confidence intervals (in black) for the set of 15 mouse lines. Red lines in each Cre graph show average values from C57BL/6J experiments. Red lines in the C57BL/6J graph are averages from the Emx1-IRES-Cre experiments, which also labels cells across all layers. There is a bias towards L5 neuron infection in both C57BL/6J and Emx1-IRES-Cre mice, highlighting the importance of using layerselective Cre lines for better coverage of cortical outputs.



Extended Data Fig. 3 | Computationally removing the distance dependence  $of connection \, weights \, alters \, the \, modular \, structure \, of the \, cortex. \, \text{To test the}$ degree to which the spatial proximity of regions affects modularity analysis, we used a power law to fit the distance component of our ipsilateral CC connectivity matrix<sup>29</sup>. Then, we repeated our modularity analysis on the 'distance-subtracted' matrix built from these residuals. a, Weighted connectivity matrix for 43 cortical areas showing the value of the residuals from a power law to fit the distance component. Rows are sources, columns are targets. Colours on the rows indicate distance-subtracted community structure with varying levels of resolution ( $\gamma = 0.5-1.5$  on the  $\gamma$ -axis,  $\gamma = 0.8$  only on the top portion of the x-axis). Columns are coloured by their module  $affiliation \, in \, the \, distance-subtracted \, matrix \, above \, their \, module \, affiliation \, in \,$ the original matrix (Fig. 1e). The inset in the top left corner shows the modularity metric (Q) for each level of  $\gamma$ , along with the Q value for a shuffled network containing the same weights. The Q values for modularity in the distance-subtracted matrix were smaller than for the original cortical matrix (for example, 0.2754 versus 0.4638 at y = 0.8) and the range of values for which

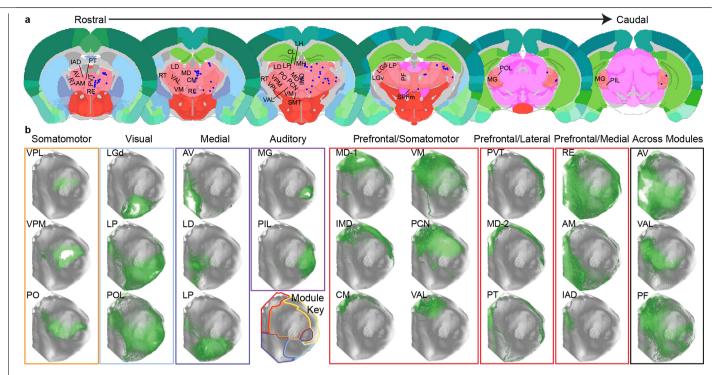
Q was greater than  $Q_{\text{shuffled}}$  was narrower (0.7  $\leq \gamma \leq$  1.7), but some modules were still present in the distance-subtracted cortical connectivity matrix. The difference between Q and  $Q_{\text{shuffled}}$  was greatest for  $\gamma = 0.8$ . The first distancesubtracted module was comprised of the entire somatomotor module, most of the lateral module, and two regions from the prefrontal module. The second distance-subtracted module contained the visual, auditory, and medial modules, plus most of the prefrontal module and one region from the lateral  $module \, (temporal \, association \, area). \, Notably, these \, modules \, were \, like \, those \,$ reported by Rubinov et al.<sup>9</sup>. As y increased past 1.0, regions began to split from the two large modules in small groups that generally did not reflect the original divisions, except for the auditory areas. b, Ipsilateral cortical network in 2D using a force-directed layout algorithm. Nodes are colour coded by module. Edge thickness shows residual values and edges between modules are coloured as a blend of the module colours. c, Cortical regions colour-coded by their distance-subtracted community affiliation at y = 0.8 show spatial relationships.



 $\textbf{Extended Data Fig. 4} \ | \ See \ next \ page \ for \ caption.$ 

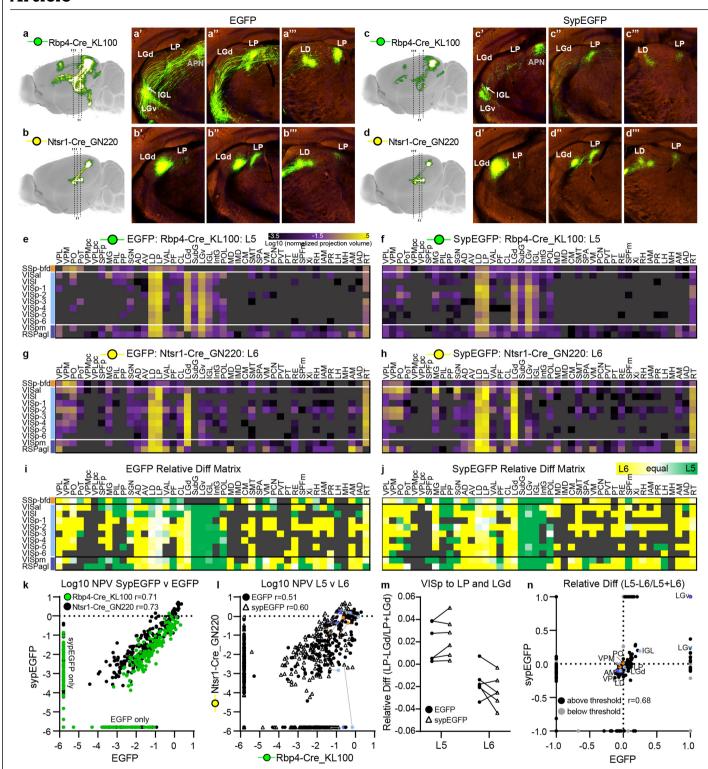
Extended Data Fig. 4 | Whole-brain single-neuron reconstructions reveal L4 IT projections. a, L4 neurons are classified into at least three morphological types as shown. b, Image shows sparse labelling of L2/3 and L4 neurons in the tamoxifen-inducible Cux2-IRES-CreERT2 driver crossed with the Ai166 reporter and using a low dose of tamoxifen via oral gavage for 1 day. L4 neurons were identified on the basis of their apical dendrite and local axons, using additional anatomical context when possible. Reconstruction was performed using Vaa3D-TeraVR on the high-resolution whole-brain image stack (composed of more than 10,000 images, resolution  $x \times y \times z$ :  $0.3 \times 0.3 \times 1 \, \mu m$ ) acquired with a two-photon fMOST system. c, We identified 25 L4 neurons for complete morphological reconstruction of dendrites and axons for three cell types and three cortical areas. In this Cre line at least, spiny stellate cells (SSCs) were most frequently identified. d, Dorsal surface view shows the CC projection patterns from three anterograde tracer experiments into the

predominantly L4 Cre lines for somatosensory cortex (SSp-m), visual cortex (VISp) and auditory cortex (AUD).  $\mathbf{e}$ - $\mathbf{k}$ , Each panel shows two examples of reconstructed cells of the same L4 type in somatosensory, visual or auditory cortex. Local morphology for each cell is shown in the inset. Arrowheads indicate axon clusters outside local region. Red, axon; blue, basal dendrite; black, apical dendrite. Consistent with canonical descriptions, we found SSCs in the somatosensory cortex that had only local axon clusters ( $\mathbf{e}$ ). However, even in these cases, we frequently observed what appeared to be an aborted axon branch (no terminal cluster found; long arrow). We also found SSCs in somatosensory cortex that did have clear axon clusters in nearby areas ( $\mathbf{g}$ ), and, in auditory cortex, SSCs projected even to the opposite hemisphere ( $\mathbf{f}$ ).  $\mathbf{h}$ - $\mathbf{k}$ , Although we identified fewer tufted pyramidal (TPC) and untufted pyramidal (UPC) cell types in this experiment, for both types we still found cells with near and long-range projections.



Extended Data Fig. 5 | Locations and cortical projection patterns from thalamic tracer experiments. a, Locations of the thalamic tracer injection centroids (blue dots) mapped onto virtual 2D coronal planes from the Allen CCFv3. To minimize the number of sections shown, all centroids are mapped within 200  $\mu$ m of their original location. See Supplementary Table 1 (thalamus tab) for more details on Cre lines and coverage. b, Example TC projections are shown in a flat map view of the ipsilateral cortical hemisphere for different thalamic nuclei arranged by the clusters identified in Fig. 3 and related to

cortical modules. Most thalamic clusters projected primarily to a single module (Fig. 3c), but some thalamic regions projected across multiple modules (for example, anteroventral nucleus (AV), ventral anterior-lateral complex (VAL), parafascicular nucleus (PF), and central lateral nucleus (CL)), or projected strongly to both prefrontal and another module; for example, somatomotor (mediodorsal nucleus (MD)-1, ventral medial nucleus (VM)), lateral (paraventricular nucleus (PVT), MD-2, parataenial nucleus (PT)) or medial regions (nucleus of reuniens (RE), anteromedial nucleus (AM)).

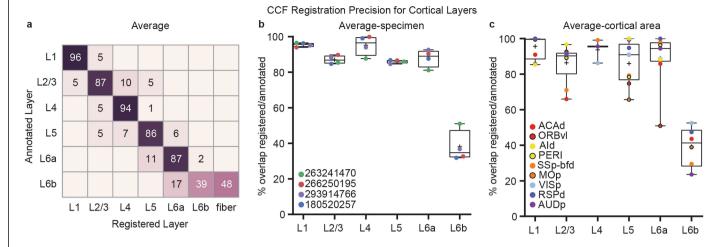


**Extended Data Fig. 6** | See next page for caption.

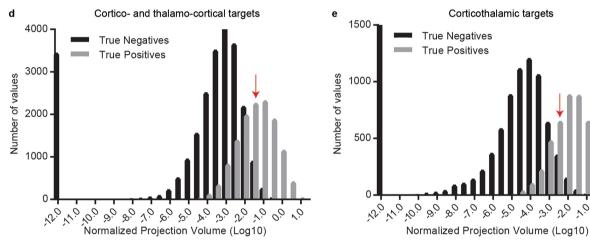
Extended Data Fig. 6 | Comparison of corticothalamic projection strengths derived from EGFP and SypEGFP tracer experiments. a-d, Maximum intensity projections from four experiments within 500 µm of each other targeting VISp (same experiment labelled VISp-3 below) using either EGFP or SypEGFP tracers in the Rbp4-Cre\_KL100 (L5) or Ntsr1\_Cre\_GN220 (L6) line as  $indicated. \textbf{\textit{a'}-d'}, Coronal\,STPT\,images\,near\,the\,centre\,of\,the\,densest\,terminal$ zone in LGd show axon and presynaptic terminal labelling in LGd and other thalamic targets, including the ventral lateral geniculate (LGd, LGv), the intergeniculate leaflet (IGL) and the lateral posterior nucleus (LP). The anterior pretectal nucleus (APN) in the midbrain is also indicated. SypEGFP labelling is more punctate and has less fluorescence in axons and fibre tracts. a"-d", Coronal STPT images near the centre of one of the densest terminal zones in the middle of LP. a'''-d'''. Coronal STPT images near the centre of the second densest terminal zone in the anterior part of LP. This image also contains a portion of the terminal zone in the lateral dorsal nucleus (LD). e-h, Directed, weighted, connectivity matrices (11 × 44) showing log<sub>10</sub>-transformed normalized projection volumes for the Cre lines representing CT projections labelled from layers  $5(\mathbf{e}, \mathbf{f})$  or  $6(\mathbf{g}, \mathbf{h})$  with EGFP or SypEGFP tracer as indicated. True negatives (including passing fibres) at the regional level were masked and

coloured dark grey. The colour map is the same as in Fig. 4. The matrix shows

relative differences for connections originating from L5 versus L6 (L5 - L6/ L5+L6) for EGFP-based measures (i) and SypEGFP-based measures (j). **k**, Normalized projection strengths for CT targets (n = 484) were significantly correlated from matched cortical locations between EGFP and SypEGFP tracers for both Cre lines (Spearman r = 0.71, 0.73; P < 0.0001). On average, EGFP CT NPVs were ~0.5 log unit larger than SypEGFP for Rbp4 experiments, but were not different for the Ntsr1 line. I, Normalized projection strengths for CT targets (n = 484) contacted by L5 or L6 cortical neurons in matched injection locations were also significantly correlated for both EGFP and SypEGFP tracers (Spearman r = 0.51, 0.60; P < 0.0001), although more weakly than for the same line between viruses. Specific connections with different fibre to terminal ratios are coloured by source module (light blue, from VISp; orange, from SSp; dark blue, from RSPagl). m, Relative differences in projection strength to LP and LGd are plotted from n = 6 VISp injection experiments (VISp-1 to VISp-6 in matrix rows above) for each Cre line and viral tracer. n, Relative difference ratios calculated for L5 to L6 using EGFP are plotted against those obtained using SypEGFP (n = 484 CT connections, n = 278 above threshold). There is a significant correlation (Spearman r = 0.68, P < 0.0001). Specific connections are coloured by source module (from I) and labelled with the target.



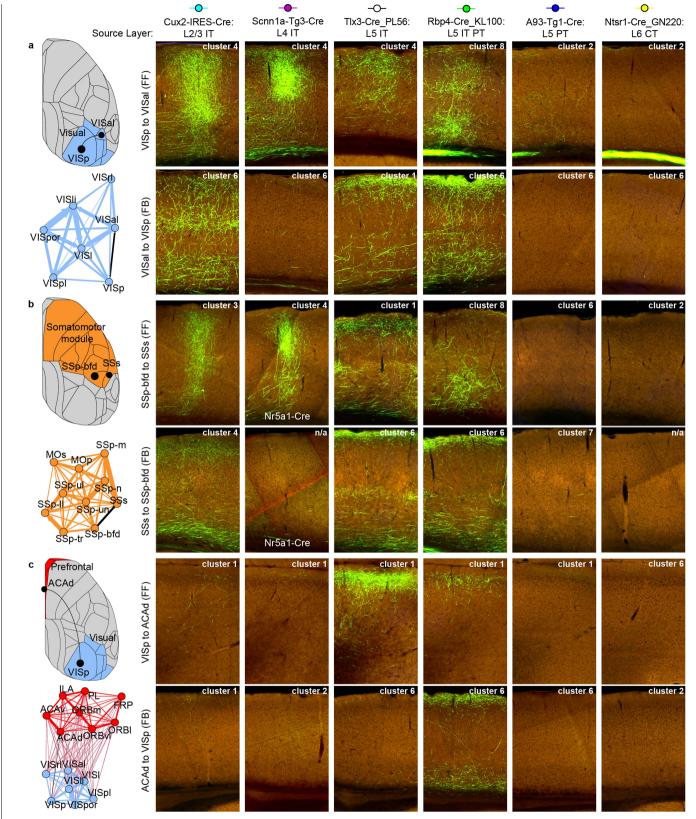
Distribution of connection weights for manually checked targets



Extended Data Fig. 7 | Validation of informatics-processing steps: CCF registration and quantification from segmentation. a-c, To determine the precision of the registration process on which we rely here for quantification of signal by layer in the cortex, we manually delineated layers 1 to 6b, using background fluorescence in coronal STPT images, for n = 9 cortical areas (ACAd, ORBvl, AId, PERI, SSp-bfd, MOp, VISp, RSPd, and AUDp; see Supplementary Table 3) in n = 4 mice per region. We then quantified the percentage of voxels within each manually annotated layer that were assigned to all cortical layers following automated registration to the CCFv3.a, A confusion matrix show the mean percentage of overlapping voxel labels averaged across these areas (individual region data in Supplementary Table 7). **b**, **c**, Boxplots show the median and mean (indicated with +); whiskers show the minimum-maximum range for the percentage overlap for individual experiments (b) or cortical areas (c, coloured dots). Across these cortical areas, the average percentage overlap ranged from 86 to 96% of voxels appropriately registered for all layers, except for L6b, which was not included in subsequent layer quantifications. For some areas and layers, the precision was worse than others; for example, while 66% of voxels were appropriately assigned to L2/3 in

ACAd, the remaining 34% were assigned to neighbouring L5. In ORBvl, only 51%of voxels were appropriately labelled for L6a. Note, however, that delineating  $layer\,5\,from\,L6a\,in\,ORBvl\,in\,coronal\,sections\,using\,just\,background$ fluorescence was very difficult even for experienced anatomists, so some of the imprecision may in fact come from the manual drawing. Even with these exceptions noted, in all cases a large majority of voxels were registered and assigned correctly. d, e, Frequency distributions of informatically derived  $quantification for manually verified true \, negative \, and \, positive \, targets. \, \boldsymbol{d}, The$ numbers of  $\log_{10}$ -transformed normalized projection values are plotted for all CC and TC targets manually verified as true negative (n = 24,272) or true positive (n = 12,921). Most true positive values were between  $\log_{10} = -4$  and  $\log_{10} = 1$ . At  $\log_{10} = -1.5$  (red arrow), 639 true negatives remained (2.6%), while 7,100 true positives were still included (54.9%), resulting in a false positive rate of 8.3% at this threshold level.  $\mathbf{e}$ , Numbers of  $\log_{10}$ -transformed normalized projection values plotted for all CC and TC targets manually verified as true negative (n = 15,789) or true positive (n = 4,503). At  $\log_{10} = -2.5$  (red arrow), 362 true negatives remained (2.3%), while 3,335 true positives were still included (74.1%), resulting in a false positive rate of 9.8% at this threshold level.

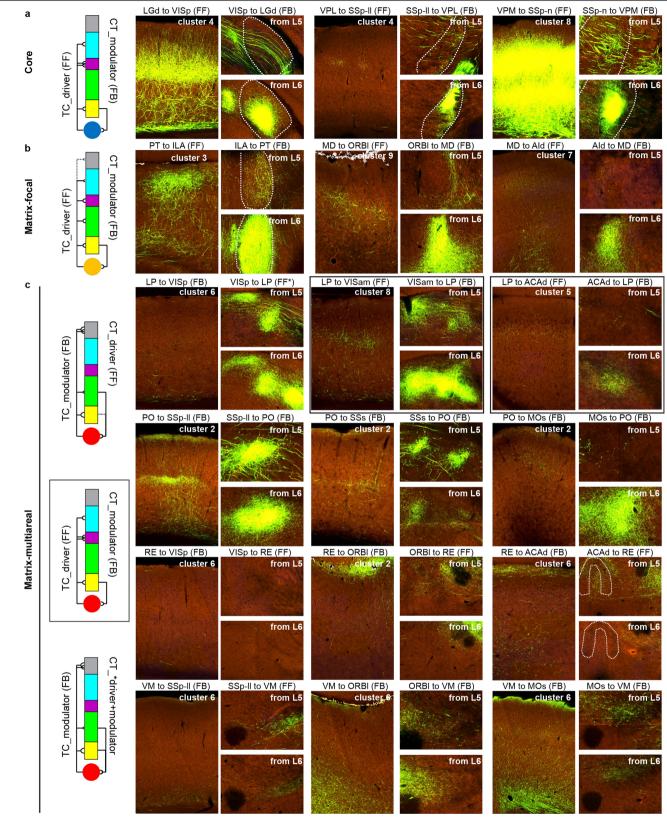
20



 $\textbf{Extended Data Fig. 8} | See \ next \ page \ for \ caption.$ 

Extended Data Fig. 8 | CC projection patterns by layer and class between reciprocally connected areas with known hierarchy. a, In the visual module, VISp and VISal (see Supplementary Table 3) are reciprocally connected (black line). VISp is the de facto bottom of visual cortex hierarchy. The output to VISal from VISp is feedforward (FF). The reciprocal connection (VISal to VISp) is feedback (FB). In the FF direction (top), VISp projections from L2/3, L4, and L5 IT projections were densest in L2/3-L5 of VISal, and relatively sparse in L1 and L6 (cluster 4). Rbp4 projections from VISp to VISal were densest in L4 and L6, with moderate levels in L2/3 (cluster 8). L5 PT and L6 CT cells projected, albeit sparsely, to L1 and L5 (cluster 2). In the FB direction (bottom), L2/3 IT axons were broadly distributed across layers, with a sparser region in L5 (cluster 6). VISal L4 IT cells projected noticeably more weakly to VISp (as opposed to the panel above), and terminated with a different pattern (L1 and L5/6, cluster 6), L5 IT cells projected densely to superficial layers in VISp (cluster 1). Rbp4 axons were dense in L1 and deep layers (cluster 6). Projections from L5 PT and L6 CT cells were also sparse, but present in L1 and L6 (cluster 6). **b**, In the somatomotor module, SSp-bfd and SSs cortex are reciprocally connected. SSpbfd to SSs is FF; the reverse is FB. In the FF direction (top), L2/3 and L4 IT cells preferentially innervate L2/3-L5, with relatively fewer terminals in L1 and L6 (clusters 3 and 4). L5 IT projections densely innervate L1 and L2/3 (cluster 1). Rbp4 projections were densest in L4 and L6, with moderate levels in L2/3

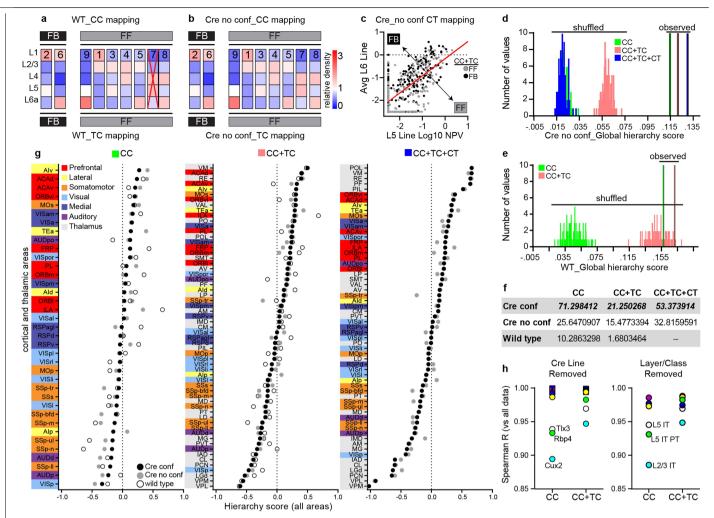
(cluster 8). L5 PT and L6 CT cell projections were sparse, and to L1 and/or deep layers (cluster 2 and 6). In the FB direction (bottom), the patterns looked remarkably like FB projections from VISal to VISp. Note again the strong connection originating from L4 cells only in the FF direction. c, VISp (in the visual module) and ACAd (in the prefrontal module) are reciprocally connected. ACAd exerts top-down control of VISp activity (FB); the reverse (VISp to ACAd) is considered FF. In the FF direction (top), L2/3, L4, and L5 cells all preferentially innervate L1 (cluster 1). In the FB direction (bottom), L2/3 cells also predominantly terminate in L1, but L5 cells project to both L1 and deep layers (L5 and L6, cluster 6). Note also there is a potentially significant sub-layer distinction; axons from VISp to ACAd are relatively deeper in L1 (or at the  $border\,of\,L1\,and\,L2/3)\,of\,ACAd, compared\,to\,the\,more\,superficial\,termination$ of ACAd axons in L1 of VISp. All panels: overall, FF projections are more often in clusters 1, 4, and 8, and FB projections in cluster 6. Cluster assignments are indicated in each panel; n/a indicates that the connection was either absent or below threshold for clustering. Areas in each module are shown in a top down cortex view and the network as a force-directed layout (edges denote normalized connection density from Fig. 1e). STPT images in the approximate centre of each target region show the laminar distribution of axons arising from labelled neurons in the different Cre lines. Images are rotated so that the pial surface is always at the top of each panel.



**Extended Data Fig. 9** | See next page for caption.

Extended Data Fig. 9 | TC and CT projection patterns and rules between reciprocally connected areas. a, Schematic summarizes observed projection patterns between core thalamic nuclei (blue circle) and their reciprocally connected cortical targets (L1-L6 colour coded). Laminar patterns are from Fig. 5g. STPT images of labelled axon terminals between three pairs of core nuclei and primary sensory cortex that perfectly follow rules in both directions. In the FF direction (LGd to VISp, VPL to SSp-II, VPM to SSp-n), projections are dense in L4 or L4 and L6 (clusters 4, 8). In the FB direction, CT projections predominantly arise from L6. b, Schematic summarizes observed projection patterns between matrix-focal thalamic nuclei (orange circle) and their reciprocally connected cortical targets. STPT images of reciprocal connections between PT and ILA, MD and ORBI, and MD and Ald illustrate the schematized rules. Projections from these thalamic nuclei belong to clusters with relatively fewer L1 axons (FF-like, clusters 3, 7, 9). The reciprocal CT input is also stronger from L6 (FB), like the core nuclei above. c, Three schematics are  $shown \,to\,summarize\,observed\,projection\,patterns\,between\,matrix-multiareal$ thalamic nuclei (red circles) and their reciprocally connected cortical targets. The top schematic shows dense TC projections to L1 (FB) with CT projections originating from L5 (FF). The middle schematic (with relevant example images boxed) shows reciprocal connection patterns in which TC projections target mid-layers (FF-like) and the reciprocal CT input is stronger from L6 (FB). The  $bottom\,schematic\,shows\,the\,same\,TC\,projection\,pattern\,as\,the\,top\,schematic,$ 

but with CT projections originating approximately equally from L5 and L6. STPT images show reciprocal connections between multiarea-matrix thalamic regions LP, PO, RE, and VM to three cortical targets each. Some regions have target-specific projections that are either FF or FB. For example, different from the LP-to-VISp projection (FB), axons from LP to VISam and ACAd target midlayers as opposed to L1 (clusters 8 and 5, FF), and the reciprocal connection  $arises\,more\,from\,L6\,(typical\,for\,FB).\,Projections\,from\,PO,RE,and\,VM\,to\,all$ three cortical targets are consistent with a FB projection (denser terminations in L1 and either L5 or L6 (clusters 2 and 6). Reciprocal CT projections originate from L5 or, both L5 and L6. We did not see CT input arising equally from both layers or more from L5 when the reciprocal TC projection was considered FF, consistent with the 'no-strong-loops' hypothesis<sup>37</sup>. All panels: overall, FF projections from core thalamic regions are in clusters 4 and 8. FB projections from matrix-multiareal thalamic regions are in clusters 2 and 6, like CC FB. The matrix-focal results support the notion that patterns with relatively less L1 involvement (3, 5, 7, 9) are FF, particularly given the strong reciprocal input observed from L6. STPT images are from the approximate centre of the axon termination field for each target region. Cortex images were rotated so that the pial surface is at the top. Cluster assignments (for TC) are indicated in each panel. Text labels above image show FF and FB direction based on relative position in Fig. 6. Dashed lines indicate region borders.



#### Extended Data Fig. 10 | Robustness of the hierarchical organization results.

We constructed multiple hierarchies using only C57BL6/J and Emx1-IRES-Cre experiments (WT) or Cre data without the Cre line confidence measure to compare with results in Fig. 6. The hierarchical position of each area  $H_i^0$  and the CC global hierarchy score  $h_{CC}$  are defined as in Eqs. (4, 5) in Methods, but with the same confidence for all lines, that is, conf(T) = 1 for all Cre lines (T). **a**, **b**, In both cases, connection types 2 and 6 are assigned to one direction (feedback), while other clusters are grouped to the opposite direction (feedforward). Cluster 7 was not identified in the WT data set. c, CT connections were also classified as in Fig. 6b for the Cre data. CT connections were not included for WT as these are exclusively defined by Crelines. d, e, Global hierarchy scores from the original, observed data, and the distributions of hierarchy scores obtained from shuffled data sets (n = 100) are shown for CC connections only (green), compared to scores obtained when TC and CT connections are sequentially included (pink, blue). The upper bound scores for an artificially perfect hierarchy using the WT data sets (e) are 0.630 for CC and 0.601 for CC + TC connections. f, z-scores were calculated for the global hierarchy scores compared to shuffled data for each of the three versions of cortical hierarchy (CC, CC+TC, CC+TC+CT). The highest z-scores were observed when using Cre line confidence weighting (compared to those with no confidence weighting or wild type data only). g, Predicted hierarchical positions of 37 cortical and 24 thalamic areas based on CC, CC+TC, or CC+TC+CT connections. Areas are ordered in each panel by the scores obtained using Cre line data with confidence weighting (Cre conf, black circles). Scores from Cre line data without confidence weighting (grey circles) and scores from wild type/Emx1-IRES-Cre data (open circles) are plotted for direct comparison. y-axis labels are colour coded by module assignment (for cortical areas). h, Robustness of the cortical hierarchy (w/ Cre conf) against individual Cre lines and projection  $classes.\,Left, Spearman\,rank\,correlation\,coefficients\,between\,the\,CC\,and$ CC + TC hierarchy with n = 13 layer- or class-specific Cre lines included versus each of the Cre lines removed. Right, results when data from Cre lines with the same layer and class were removed together. Removal of these lines and classes produced relatively minor deviations from the overall hierarchy determined with all data. Note that in both panels the y-axis starts at r = 0.85. For all lines and classes, the correlation with the hierarchy using the complete data set is very high. The lowest correlations occurred following removal of Cux2-IRES-Cre, Rbp4-Cre\_KL100, and Tlx3-Cre\_PL56.



# **Reporting Summary**

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see <u>Authors & Referees</u> and the <u>Editorial Policy Checklist</u>.

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main

# Statistical parameters

text,	ext, or Methods section).					
n/a	Cor	nfirmed				
	$\boxtimes$	The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement				
	$\boxtimes$	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly				
	$\boxtimes$	The statistical test(s) used AND whether they are one- or two-sided Only common tests should be described solely by name; describe more complex techniques in the Methods section.				
	$\boxtimes$	A description of all covariates tested				
	$\boxtimes$	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons				
	$\boxtimes$	A full description of the statistics including <u>central tendency</u> (e.g. means) or other basic estimates (e.g. regression coefficient) AND <u>variation</u> (e.g. standard deviation) or associated <u>estimates of uncertainty</u> (e.g. confidence intervals)				
$\boxtimes$		For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i> ) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted <i>Give P values as exact values whenever suitable.</i>				
$\times$		For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings				
	$\boxtimes$	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes				
		Estimates of effect sizes (e.g. Cohen's d. Pearson's r), indicating how they were calculated				

Our web collection on <u>statistics for biologists</u> may be useful.

## Software and code

Policy information about availability of computer code

State explicitly what error bars represent (e.g. SD, SE, CI)

Clearly defined error bars

Data collection

Serial 2 photon images were processed using the Allen informatics data pipeline (IDP), which manages the processing and organization of the images and quantified data for analysis and display in the web application as previously described (Oh et al., 2014 and Kuan et al. 2015). Illumination and image acquisition for intrinsic signal imaging were controlled with an in-house GUI software written in Python. For single cell morphologies, following acquisition of the complete fMOST image stack, it was converted to a multi-level navigable dataset using the open source Vaa3D-TeraFly program then reconstructions were performed using Vaa3D-TeraVR software tools built to facilitate semi-automated and manual reconstructions (Bria et al., 2016 and Wang et al., 2019).

Data analysis

Unsupervised hierarchical clustering was conducted with the online software, Morpheus, (https://software.broadinstitute.org/morpheus/) for algorithms and for visualization of the dendrogram and heat maps. The software program GraphPad Prism was used for statistical tests and generation of all graphs, and the software program Gephi was used for visualization and layout of network diagrams. The software program Vaa3D was used for visualization of single cell morphologies. Hierarchy analyses were performed as described in detail in Methods. Code and data files for hierarchical analyses are available through the Allen SDK and Github (https://github.com/AllenInstitute/MouseBrainHierarchy).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data (including high resolution images, segmentation, registration to CCFv3, and automated quantification of injection size, location, and distribution across brain structures) are available through the Allen Mouse Brain Connectivity Atlas portal (http://connectivity.brain-map.org/). Individual experiment summaries can be viewed using this link:http://connectivity.brain-map.org/projection/experiment/[insert experimental id]. In addition to visualization and search tools available at this site, users can download data using the Allen Brain Atlas API (http://help.brain-map.org/display/mouseconnectivity/API) and the Allen Brain Atlas Software Development Kit (SDK: http://alleninstitute.github.io/AllenSDK/connectivity.html). Through the SDK, structure and voxel-level projection data are available for download. Examples of code for common data requests are provided as part of the Mouse Connectivity Jupyter notebook to help users get started with their analyses. Our code for hierarchical analyses is also available through the Allen SDK and Github (https://github.com/AllenInstitute/MouseBrainHierarchy).

Field-spe	ecific reporting					
Please select the b	est fit for your research. If you are not sure, read the appropriate sections before making your selection.					
Life sciences	Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences					
For a reference copy of	the document with all sections, see <u>nature.com/authors/policies/ReportingSummary-flat.pdf</u>					
Life scier	nces study design					
All studies must dis	sclose on these points even when the disclosure is negative.					
Sample size	In our previously published study (Oh et al., 2014), we demonstrated that an n=1 is a good predictor of connectivity strengths across multiple animals. Here, we also found that the correlations between brain-wide projection strengths from experiments at matched locations are positive and significant (r>0.8, Extended Data Figure 1). Thus, we are able to confidently and comprehensively sample across the entire cortex (and entire brain) with n=1 experiment per source area and Cre line. This consistency is what allows us to map the connectome, which still required 1,000 experiments (i.e. mice) for the coverage we determined was necessary for close to completeness.					
Data exclusions	Image series for each tracer experiment were curated for inclusion based on pre-established QC metrics, and quantitative data had to pass the threshold criteria described in the Results.					
Replication	As stated above, in our previously published study (Oh et al., 2014), we demonstrated that an n=1 is a good predictor of connectivity strengths across multiple animals. Here, we also found that the correlations between brain-wide projection strengths from experiments at matched locations are positive and significant (r>0.8, Extended Data Figure 1). Thus, we are able to confidently and comprehensively sample across the entire cortex (and entire brain) with n=1 experiment per source area and Cre line. This consistency is what allows us to map the connectome, which still required >1,000 experiments (i.e. mice) for the coverage we determined was necessary for close to completeness.					
Randomization	Randomization of animals to different groups is not relevant to our study design. We did not have experimental vs. control groups.					
Blinding	Image data acquisition and quantitative measures of projection strengths are automated, so blinding was not necessary. Investigators performing manual target analyses were not blinded to injection source or cre line, but this was impossible as the annotation required an anatomy expert to look through every single image. The location would be thus known, and the layer of origin of the cells thus obvious as well.					

# Reporting for specific materials, systems and methods

Mat	terials & experimental systems	Methods	
n/a	Involved in the study	n/a	Involved in the study
	☐ Unique biological materials	$\boxtimes$	ChIP-seq
$\boxtimes$	Antibodies	$\boxtimes$	Flow cytometry
$\boxtimes$	Eukaryotic cell lines	$\boxtimes$	MRI-based neuroimaging
$\boxtimes$	Palaeontology		
	Animals and other organisms		
$\boxtimes$	Human research participants		

# Unique biological materials

Policy information about <u>availability of materials</u>

Obtaining unique materials

Viral tracers are available through Addgene, the Penn Vector Core, or via request to the Allen Institute. Cre driver lines are available from the repositories indicated in Supplementary Table 1.

# Animals and other organisms

Policy information about <u>studies involving animals</u>; <u>ARRIVE guidelines</u> recommended for reporting animal research

Laboratory animals Mus musculus, C57Bl/6J, males and females, Cre driver transgenics, P56 (+/-7 days)

Wild animals This study did not involve wild animals.

Field-collected samples This study did not involve field-collected samples.

# Allele-selective lowering of mutant HTT protein by HTT-LC3 linker compounds

https://doi.org/10.1038/s41586-019-1722-1

Received: 5 February 2019

Accepted: 24 September 2019

Published online: 30 October 2019

Zhaoyang Li<sup>1,9</sup>, Cen Wang<sup>1,9</sup>, Ziying Wang<sup>1,9</sup>, Chenggang Zhu<sup>2,9</sup>, Jie Li<sup>3</sup>, Tian Sha<sup>1</sup>, Lixiang Ma<sup>4</sup>, Chao Gao<sup>5</sup>, Yi Yang<sup>6</sup>, Yimin Sun<sup>1</sup>, Jian Wang<sup>1</sup>, Xiaoli Sun<sup>1</sup>, Chenqi Lu<sup>1</sup>, Marian Difiglia<sup>7</sup>, Yanai Mei<sup>1</sup>, Chen Ding<sup>1,10</sup>, Shouqing Luo<sup>6,10</sup>, Yongjun Dang<sup>8</sup>, Yu Ding<sup>1\*</sup>, Yiyan Fei<sup>2\*</sup> & Boxun Lu<sup>1\*</sup>

Accumulation of mutant proteins is a major cause of many diseases (collectively called proteopathies), and lowering the level of these proteins can be useful for treatment of these diseases. We hypothesized that compounds that interact with both the autophagosome protein microtubule-associated protein 1A/1B light chain 3 (LC3)<sup>1</sup> and the disease-causing protein may target the latter for autophagic clearance. Mutant huntingtin protein (mHTT) contains an expanded polyglutamine (polyQ) tract and causes Huntington's disease, an incurable neurodegenerative disorder<sup>2</sup>. Here, using small-molecule-microarray-based screening, we identified four compounds that interact with both LC3 and mHTT, but not with the wild-type HTT protein. Some of these compounds targeted mHTT to autophagosomes, reduced mHTT levels in an allele-selective manner, and rescued disease-relevant phenotypes in cells and in vivo in fly and mouse models of Huntington's disease. We further show that these compounds interact with the expanded polyQ stretch and could lower the level of mutant ataxin-3 (ATXN3), another disease-causing protein with an expanded polyQ tract<sup>3</sup>. This study presents candidate compounds for lowering mHTT and potentially other disease-causing proteins with polyQ expansions, demonstrating the concept of lowering levels of disease-causing proteins using autophagosometethering compounds.

Lowering the levels of disease-causing proteins, especially those with unknown activities, is an emerging approach for disease treatment. Biological tools such as RNA-mediated inhibition (RNAi) or CRISPR may achieve this goal<sup>4-6</sup>, but their clinical delivery is challenging. Enhancing proteasomal degradation of target proteins using proteolysis-targeting chimeric molecules (PROTACs) is a promising emerging approach<sup>7</sup>, but proteasomes alone are inefficient in degrading certain large proteins or aggregates8. Macroautophagy (hereafter referred to as autophagy), an independent protein-degradation pathway, is a bulk degradation system that engulfs proteins into autophagosomes for subsequent lysosomal degradation<sup>9</sup>. Autophagy is present in all eukaryotic cells, and therefore harnessing the power of autophagy to degrade certain target proteins may have potential for drug discovery. Here we investigate this possibility in the context of lowering mHTT, which contains a polyQ stretch with at least 36 glutamine residues and causes Huntington's disease, an incurable monogenetic neurodegenerative disorder<sup>2</sup>.

mHTT could be degraded by autophagy, during which protein substrates are incorporated into double-membrane autophagosomes associated with lipidated LC3<sup>1</sup>. We therefore hypothesized that linker

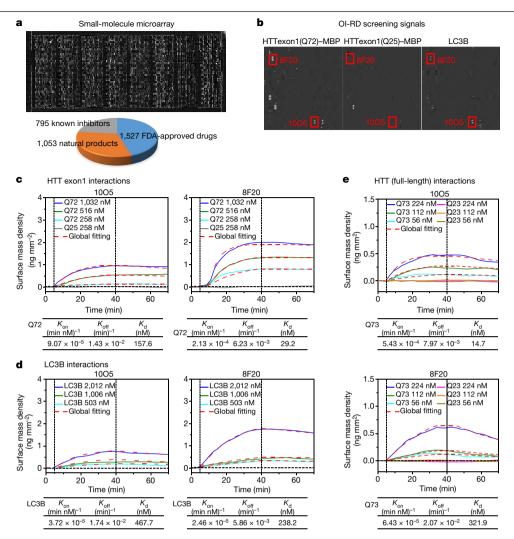
compounds that interact with both mHTT and LC3 may tether the molecules together to enhance recruitment of mHTT into autophagosomes, facilitating its degradation. In addition, mHTT–LC3 linker compounds that do not interact with wild-type HTT (wtHTT) may promote allele-selective degradation of mHTT. Because no mHTT–LC3-interacting compounds have been reported, we performed small-molecule-microarray (SMM)-based screening for such compounds and used wtHTT for the counter-screen to identify allele-selective candidates.

#### Results

#### Identification of mHTT-LC3 linker compounds

We stamped 3,375 compounds (Fig. 1a) in duplicate onto a microarray on isocyanate-functionalized glass slides using the nucleophile-isocyanate reaction, which forms covalent bonds between the compounds and the glass slides <sup>10,11</sup>. We then purified the human LC3B protein (Extended Data Fig. 1a, b, Supplementary Table 1), a pathogenic mHTT exon1 fragment <sup>12</sup> with an expanded polyQ region containing 72 glutamines (mHTTexon1(Q72)), and a control wtHTT exon1 fragment

Neurology Department at Huashan Hospital, State Key Laboratory of Medical Neurobiology and MOE Frontiers Center for Brain Science, Institutes of Brain Science, School of Life Sciences, Fudan University, Shanghai, China. Department of Optical Science and Engineering, Shanghai Engineering Research Center of Ultra-Precision Optical Manufacturing, Key Laboratory of Micro and Nano Photonic Structures (Ministry of Education), Fudan University, Shanghai, China. Brace Preparation System, National Facility for Protein Science in Shanghai, Shanghai, China. Brace Preparation System, National Facility for Protein Science in Shanghai, Shanghai, China. Peninsula Schools of Medicine and Dentistry, Institute of Translational and Stratified Medicine, University of Plymouth, UK. Laboratory of Cellular Neurobiology, Department of Neurology, Massachusetts General Hospital, Charlestown, MA, USA. Key Laboratory of Metabolism and Molecular Medicine, Ministry of Education, Department of Biochemistry and Molecular Biology, School of Basic Medical Sciences, Fudan University, Shanghai, China. These authors contributed equalty: Zhaoyang Li, Cen Wang, Ziying Wang, Chenggang Zhu. These authors jointly supervised this work: Chen Ding, Shouqing Luo. Permail: yuding@fudan.edu.cn; fyy@fudan.edu.cn; luboxun@fudan.edu.cn



 $\label{limited} \textbf{Fig. 1} | \textbf{Identification of potential mHTT-LC3 linker compounds by SMM-based screening and validation. a}, Ol-RD image of a SMM. Each compound was printed in duplicate in adjacent vertical positions. b, Magnified view of surface mass-density changes after incubation with HTTexon1(Q25)-MBP, HTTexon1(Q72)-MBP or LC3B. The red outlines highlight two hits (10O5 and 8F20). \textbf{c-e}, Association-dissociation curves of surface-immobilized compounds 8F20 and 10O5 with HTTexon1(Q72)-MBP (Q72) (\textbf{c}), LC3B (\textbf{d}) and labels are compounds of the surface of the$ 

full-length HTT(Q73) (Q73) (e) at the indicated purified protein concentrations. In association—dissociation curves, vertical dashed lines mark the starts of association and dissociation phases of the binding event. The red dashed curves are global fits to a Langmuir reaction model with the fitting parameters listed at the bottom of each plot. No binding signals were observed for HTTexon1(Q25)—MBP or full-length HTT(Q23) proteins, and thus these parameters are not presented.

(HTTexon1(Q25)) (Extended Data Fig. 1c, d) for the screen. We fused a maltose-binding protein (MBP) tag to both HTT exon1 proteins to increase their solubility for subsequent experiments.

To identify compounds that interact with LC3B and mHTT, we incubated these proteins over the SMMs and detected compoundprotein interactions using a scanning oblique-incidence reflectivity difference (OI-RD) microscope<sup>13-18</sup>, an optical biosensor. We then performed experiments with HTTexon1(Q25) or buffer alone to exclude nonspecific signals, and identified two compounds, 1005 (GW5074, 3-3-((3,5-dibromo-4-hydroxyphenyl)methylidene)-5-iodo-1*H*-indol-2-one) and 8F20 (ispinesib, N-(3-aminopropyl)-N-((1R)-1-(7-chloro-4-oxo-3-(phenylmethyl)-2-quinazolinyl)-2-methylpropyl)-4-methylbenzamide), that interact with both LC3B and mHTTexon1(Q72), but not with HTTexon1(Q25) (Fig. 1b, annotation based on the ID in the compound library). We then measured the on and off rates ( $K_{on}$  and  $K_{off}$ , respectively) of these interactions to confirm our observation (Fig. 1c, d), finding that both compounds showed dissociation constants  $(K_d)$  of around 100 nM with LC3B or mHTTexon1(Q72). As shown in Fig. 1e, these compounds also interacted with the full-length mHTT (flHTT(Q73), Extended Data Fig. 1e), but not with wild-type HTT (HTTexon1(Q25)

or flHTT(Q23), Fig. 1c, e) or irrelevant proteins (Extended Data Fig. 2a) including MBP-His $_8$  (MBP), superfolder GFP (sfGFP) and Rpn10 (a proteasomal ubiquitin receptor) (Extended Data Fig. 1f). We then validated the interaction using an orthogonal assay, microscale thermophoresis (MST), and obtained consistent results (Extended Data Fig. 2b).

#### **Linkers induced allele-selective mHTT lowering**

We then tested whether these potential mHTT–LC3 linker compounds decrease mHTT levels via autophagy as predicted. Both compounds decreased levels of mHTT in cultured primary cortical neurons from a well-established HD-knock-in mouse model  $(Hdh^{Q7/Q140})^{19}$  (Fig. 2a), but had little or no effect on levels of wtHTT in the heterozygous HD neurons  $(Hdh^{Q7/Q140})$  (Fig. 2a) or wild-type neurons  $(Hdh^{Q7/Q7})$  (Fig. 2b), consistent with the lack of interaction of these compounds with wtHTT. We then screened for other mHTT–LC3 linker compounds on the basis of common features of the two hit compounds 10O5 and 8F20. The hydroxyl group in 10O5 and the amino group in 8F20 were used in the nucleophile-isocyanate reaction for stamping of the SMMs, and these groups were inaccessible to mHTT and LC3B for the compound–protein

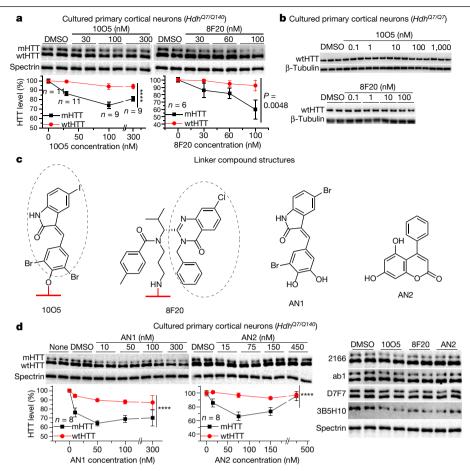


Fig. 2 | mHTT-LC3 linker compounds lower mHTT but not wtHTT in cultured mouse neurons via autophagy. a, Western blot (HTT detected by the 2166 antibody) and quantification of compound-treated cultured cortical neurons from  $Hdh^{Q7/Q140}$  HD-knock-in mice. Two-way ANOVA. For 1005, F(1,72) = 50.93, P < 0.0001; for 8F20, F(1, 40) = 8.903, P = 0.0048. **b**, Representative western blots (from three biological repeats) of cultured wild-type cortical neurons treated with the indicated compounds. c, Two-dimensional structures of the hit compounds and the other identified effective linker compounds. The red

lines indicate the glass chip surface on which the compounds are immobilized. The dotted ovals indicate the possible chemical groups exposed for proteincompound interactions in the screening. d, Left and middle, as in a, but using AN1 or AN2. For AN1, F(1,70) = 32.96, P < 0.0001; for AN2, F(1,69) = 23.03, P<0.0001. Right, as in a, but blotted with indicated HTT antibodies (1005 and 8F20:100 nM; AN2:50 nM). For all panels, n indicates the number of independently plated wells; data are mean + s.e.m. Full blots of cropped gels are shown in Extended Data Fig. 3b or Supplementary Fig. 1.

interaction during the screening (Fig. 2c). Thus, whereas the two hit compounds have different structures, the exposed chemical groups on the SMMs share similarities in that they contain an aryl ring connected to a lactam-based bicyclic structure with halogen-substituted aryl group (Fig. 2c). We tested several compounds with similar features and identified two additional mHTT-LC3 linker compounds (Fig. 2c, AN1 (3-5-bromo-3-((3-bromo-4,5-dihydroxyphenyl)methylidene)-1*H*-indol-2-one) and AN2 (5,7-dihydroxy-4-phenylcoumarin), which interact with both mHTT and LC3B but not with wtHTT or irrelevant control proteins (Extended Data Fig. 2c, d). They also reduced the levels of mHTT in an allele-selective manner in cultured HD mouse neurons (Fig. 2d). No cytotoxicity was observed in cultured neurons treated with these compounds at the tested concentration range (Extended Data Fig. 2e), confirming that the reduction in mHTT was not due to cell loss.

Most of these compounds showed an optimal dose (hook effect) in lowering mHTT (Fig. 2a, d): a sufficient concentration is desired for tethering mHTT and LC3 together, but excessively high concentrations may cause the compound molecules to interact with mHTT and LC3 separately without tethering them. Similar concentration-dependent effects were observed in fibroblasts of patients with HD (Fig. 3c, right) and have been reported for PROTAC<sup>20</sup>. Consistent with the prediction that the reduction in mHTT is mediated by degradation via autophagy, the autophagy inhibitor NH<sub>4</sub>Cl or chloroquine blocked the mHTT-lowering effects (Extended Data Fig. 3a), confirming that the compounds targeted mHTT for autophagic degradation. Further, the compoundinduced mHTT-lowering effects were only slightly enhanced by the mTOR inhibitor rapamycin, an enhancer of autophagosome formation (Extended Data Fig. 3a, right; also see Fig. 3b).

The reduction of mHTT levels could be detected by multiple mHTT antibodies-including 3B5H10, which detects a toxic species of the expanded polyQ stretch<sup>21,22</sup> (Fig. 2d, right)—suggesting that the detected reduction of the mHTT signal was not due to changes in affinity to a specific antibody. In addition, we did not observe any obvious increase of possible polyQ-containing mHTT fragments at lower molecular weights (Extended Data Fig. 3b, c), suggesting that the reduced levels of mHTT were not a result of increased site-specific cleavages of mHTT.

We further investigated the effects of the compounds in cells from patients with HD using the well-established homologous time-resolved fluorescence (HTRF) assay<sup>23,24</sup>, which is more quantitative than western blots but is not applicable to mouse mHTT proteins owing to non-specific signals<sup>25</sup>. We observed autophagy-dependent lowering of mHTT by these compounds in fibroblasts from patients with HD and neurons derived from induced pluripotent stem cells (iPS cells) (Fig. 3a, b, Extended Data Fig. 3d), but no lowering of wtHTT in fibroblasts from healthy human donors or patients with Parkinson's disease

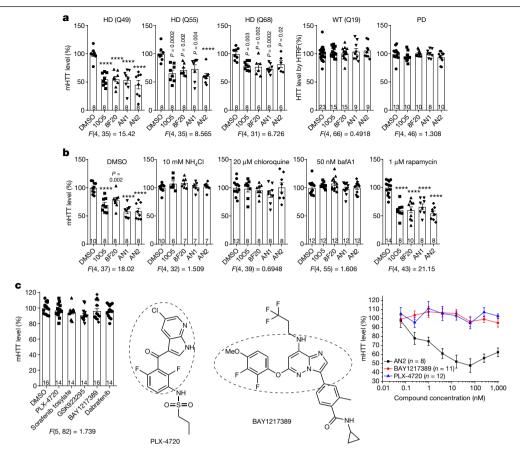


Fig. 3 | mHTT-LC3 linker compounds lower mHTT in cells from patients with HD. a, HTT levels measured by HTRF (2B7/MW1 for mHTT, and 2B7/2166 for total HTT) in primary fibroblasts from patients with HD, wild-type controls (WT) or patients with Parkinson's disease (PD) who were treated with the indicated compounds (100 nM). All signals were normalized to the average signals from the DMSO control group. One-way ANOVA with post hoc Dunnett's tests. \*\*\*\*P<0.0001. b, As in a, but using immortalized fibroblasts treated with or without the autophagy inhibitors NH<sub>4</sub>Cl, chloroquine or bafA1, or the

autophagy enhancer rapamycin.  $\mathbf{c}$ , Left, as in  $\mathbf{a}$ , but using immortalized fibroblasts from patients with HD (expressing mHTT(Q47)), treated with indicated c-Raf or KSP inhibitors at 100 nM. Middle, 2D structure of the inhibitors. The dotted ovals indicate the parts of the compounds that share similarities with the hit compounds. Right, dose-response curves of the indicated compounds. For all panels, n indicates the number of independently plated wells; data are mean  $\pm$  s.e.m.

(Fig. 3a). To further confirm the role of autophagic degradation, we tested the effects of the compounds with or without lowering of ATGS, a key autophagy gene that is required for autophagosome formation  $^{26}$ . ATG5 knockdown in fibroblasts from a patient with HD (expressing mHTT(Q47)) significantly decreased LC3-II levels and nullified the mHTT-lowering effects induced by the mHTT-LC3 linker compounds (Extended Data Fig. 3e). Similar results were obtained in ATG5-knockout mouse embryonic fibroblasts  $^{26}$  (MEFs, Extended Data Fig. 3f), confirming that the effects of the compounds were mediated by autophagic degradation.

The two hit compounds 1005 and 8F20 are known to inhibit c-Raf and KSP<sup>27,28</sup>, respectively, whereas AN1 and AN2 had unknown activities on these targets. We therefore tested their potential influence on c-Raf and KSP. On the basis of the in vitro c-Raf kinase assay, 1005 (a known c-Raf inhibitor)—but not the other three compounds—inhibited c-Raf at the concentrations tested (Extended Data Fig. 4a). We then tested MEK and ERK phosphorylation levels in the cultured neurons treated with these compounds at optimal mHTT-lowering concentrations to evaluate Raf activity<sup>29</sup> and found no significant effects of all tested compounds (Extended Data Fig. 4b, left). We also tested phospho-BUBR1 levels to evaluate KSP activity<sup>30</sup>, and again observed no significant effects (Extended Data Fig. 4b, right). We made similar observations in fibroblasts (expressing mHTT(Q47)) from a patient with HD (Extended Data Fig. 4c). Thus, the observed reduction in mHTT is probably irrelevant

to c-Raf or KSP inhibition. To further confirm this, we examined the effects of several known c-Raf or KSP inhibitors, and found that they had no HTT-lowering effects (Fig. 3c, left). Two of these inhibitors, PLX-4720 and BAY1217389, have structures similar to 1005 and 8F20, respectively (Fig. 3c, middle). These compounds did not lower mHTT in cells from patients at sub-micromolar concentrations (Fig. 3c, right), probably because they had very weak affinity to LC3 and mHTT, if any (Extended Data Fig. 2c, right). By contrast, AN2 reduced mHTT levels in the same cells in a dose-dependent manner (Fig. 3c, right).

We then investigated the effects of the compounds in vivo. Because the *Drosophila* LC3 homologue Atg8 has a predicted structure that is highly similar to LC3B (Extended Data Fig. 5a), we tested the compounds in a HD transgenic fly model expressing human full-length mHTT. All of the mHTT–LC3 linker compounds that we identified significantly reduced mHTT levels in *Drosophila* (Extended Data Fig. 5b), validating the in vivo efficacy of these compounds.

We further investigated the in vivo effects of the compounds using the HD-knock-in mouse model ( $Hdh^{Q7/Q140}$ )<sup>19</sup> by intracerebroventricular injections. Treatment with three of the four linker compounds (1005, AN1 or AN2, but not 8F20) led to significant lowering of mHTT in cortices of HD mice (Extended Data Fig. 6a). We then performed intraperitoneal injection of 10O5 and AN2 at 0.5 mg kg $^{-1}$  in HD-knock-in mice. The compounds crossed the blood–brain barrier and reached the brain at detectable concentrations (Extended Data Fig. 5c, approximately

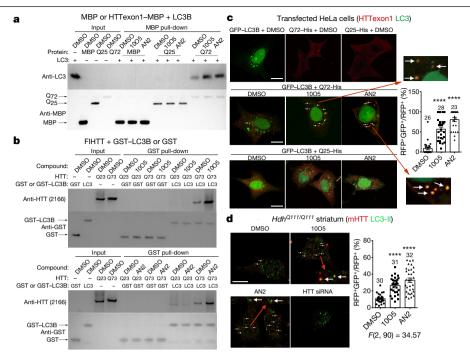


Fig. 4 | Linker compounds enhance the mHTT-LC3 interaction and tether  $\textbf{mHTT to autophagosomes.} \textbf{\textit{a}}, \textbf{\textit{b}}, \textbf{Representative results} (from three biological and \textbf{\textit{c}}) and \textbf{\textit{c}}) are the properties of the properti$ repeats) of in vitro pull-down experiments using purified HTT and LC3B proteins (see Methods for details). c, d, Representative images (scale bar,

10 µm) and quantification of the co-localization between HTT and autophagosomes in HeLa cells transfected with the indicated cDNA plasmids (c) or striatum from  $Hdh^{QIII/QIII}$  mice (d). Data are mean  $\pm$  s.e.m. n indicates the number of cells. One-way ANOVA with post hoc Dunnett's test. \*\*\*\*P < 0.0001.

20-200 nM for 1005 and 20-40 nM for AN2; no signal was detected in the DMSO-injected control group) 0.5-6 h after injection. Consistent with these results, we observed significant allele-selective lowering of mHTT in mouse cortices and striata (Extended Data Fig. 6b, c). The observed lowering was not due to changes in mHTT solubility, because no increase of mHTT aggregates was observed in the cortical tissues of mice treated with these compounds (Extended Data Fig. 6d).

#### Linkers tether mHTT to autophagosomes

We then examined whether these compounds actually function as linkers between mHTT and LC3 to target mHTT for autophagosome engulfment. The presence of 1005 or AN2, the two compounds that were effective by intraperitoneal injection in vivo, markedly enhanced the mHTT-LC3 interaction in in vitro pull-down experiments (Fig. 4a, comparing lane 11 with 12 and 13; Fig. 4b, comparing lane 11 with 12, top and bottom). There was no enhancement effect for the wild-type HTT-LC3 interaction (Fig. 4a, lanes 8-10; Fig. 4b, lane 9 and 10, top and bottom). Consistent with this, these compounds led to increased engulfment of mHTT by autophagosomes, both in transiently transfected HeLa cells expressing exogenous GFP-LC3B and in HTTexon1-MBP-His fragments (Fig. 4c), and in mouse striatal cells (STHdh<sup>Q111/Q111</sup>)<sup>31</sup> expressing endogenous LC3 and full-length mHTT proteins (Fig. 4d).

These data confirmed that the compounds tether mHTT, LC3B and autophagosomes in vitro and in cells, although the detailed structural information remains to be resolved.

#### Linkers do not influence autophagy function

The lowering of mHTT levels by the linker compounds was unlikely to be a result of enhanced autophagy, because the number and size of autophagosomes remained unchanged (Extended Data Fig. 7a). We further investigated whether the compounds could influence autophagy using established approaches<sup>32-34</sup>. Neither 1005 nor AN2 influenced the autophagosome-lysosome fusion or autophagy activity (Extended Data Fig. 7b-d). Furthermore, we observed no changes in LC3-II levels in the cultured cortical neurons treated with 1005 or AN2 in the absence or presence of the lysosome inhibitor bafilomycin A1 (bafA1) (Extended Data Fig. 7e). The level of the known autophagy-selective substrate protein SQSTM1 (also known as p62) was also unaffected in vivo and in cultured neurons (Extended Data Figs. 7f, 8a). In addition, other wild-type polyQ proteins (ATXN3 and TBP) and control proteins (NBR1, NCOA4, actin, GAPDH and tubulin) were not influenced (that is, any change amounted to less than 10%) (Extended Data Fig. 8a).

We then performed proteomics analysis to obtain a more complete overview of proteins that may have been influenced by these compounds. We observed significant lowering (about 20%, P < 0.01) of HTT levels in cortices of mice injected intraperitoneally with 1005 or AN2 (Extended Data Fig. 8b, bar plots). As the proteomics analysis was unable to distinguish mHTT from wtHTT, the actual reduction in mHTT is likely to be higher than this. Meanwhile, using the criteria of P < 0.01, we observed changes in only a small percentage of proteins (Extended Data Fig. 8b; see Supplementary Table 2 for details). No autophagyspecific substrate proteins exhibited significant changes and there was no enrichment of proteins associated with the autophagy pathway (Supplementary Table 2), further confirming that autophagy was unaffected. Proteomics analysis in cultured neurons gave consistent results (Extended Data Fig. 8c; see Supplementary Table 3 for details).

#### Linker compounds depleted expanded polyQ proteins

The linker compounds interacted with and lowered mHTT but not wtHTT (Fig. 1). The simplest explanation for this specificity is that the compounds specifically interact with the expanded polyQ tract, possibly by recognizing its emergent conformation, which is different from that of the short polyQ stretch<sup>21,35</sup>. If so, the linker compounds may also affect other proteins with expanded polyQ regions. Consistent with this prediction, compounds 1005, AN1 and AN2 reduced the levels of mutant but not wild-type ATXN3 in fibroblasts from patients with spinocerebellar ataxia type 3 (SCA3) (Extended Data Fig. 9a) and

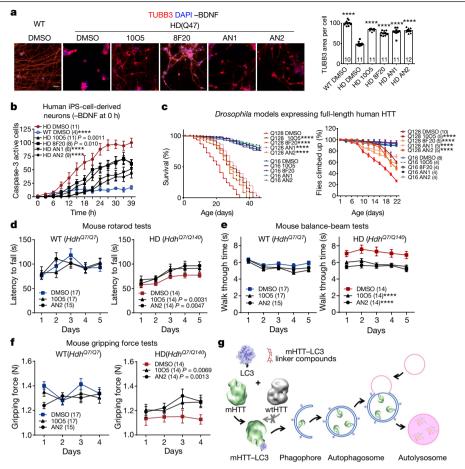


Fig. 5 | Linker compounds rescue HD-relevant phenotypes in cells and in vivo. a, Left, representative DAPI staining and immunostaining of the neuronal-specific tubulin marker TUBB3, showing neuronal morphology of patient iPS-cell-derived striatal neurons (HD: Q47; WT: Q19) treated with indicated compounds. Right, quantification of TUBB3 area per cell. Scale bar,  $50\,\mu\text{m}$ . b, Neuronal apoptosis at different time points after BDNF removal, measured using a green fluorescent dye (NucView 488) to detect active caspase-3. c, Left, Kaplan–Meier survival curves of transgenic *Drosophila* with

the indicated transgenes and treatments. Right, climbing performance of the treated transgenic flies as a function of age after eclosion.  $\mathbf{d}$ – $\mathbf{f}$ , Mouse behavioural tests showing improvement of HD-relevant phenotypes after intraperitoneal injection of the indicated compounds at 0.5 mg kg $^{-1}$ .  $\mathbf{g}$ , A model showing how mHTT–LC3 linker compounds may induce mHTT degradation, illustrating the concept of lowering target protein levels using autophagosome-tethering compounds. The images representing HTT are reproduced from ref.  $^{37}$  under a CC BY license.

exogenously expressed 72Q-GFP, 46Q-GFP and 38Q-GFP but not 25Q-GFP proteins (containing on Met-polyQ-sfGFP sequences) in HEK293T cells (Extended Data Fig. 9b). These data suggest that the compounds distinguished the expanded polyQ stretch from the short polyQ stretch at a threshold between 25Q and 38Q. To further confirm this, we tested the interactions of these compounds with polyQ motifs (Extended Data Fig. 9c) and confirmed that 10O5, AN1 and AN2 interact with polyQ-GFP with 38 or more glutamine residues, but not 25Q-GFP or GFP alone (Extended Data Figs. 2, 9d, e).

#### Linker compounds rescued HD-relevant phenotypes

We further investigated the therapeutic potential of the compounds for treating HD. All the mHTT–LC3 linker compounds rescued mHTT toxicity in neurons derived from iPS cells of patients with HD (Fig. 5a, b). They also rescued HD-relevant behavioural deficits and increased the lifespan of flies expressing human mHTT, while having no influence on the flies expressing wtHTT (Fig. 5c).

Finally, we investigated the disease-relevant behavioural phenotypes in ten-month-old heterozygous HD-knock-in mouse ( $Hdh^{Q7/Q140}$ ). HD mice exhibited significant deficits in several behavioural tests, including rotarod, balance beam and gripping force tests (Extended Data Fig. 9f–h). Intraperitoneal injection of 1005 or AN2, but not of DMSO

only, significantly improved HD-relevant behavioural deficits in these tests, without influencing the wild-type mice (Fig. 5d–f), demonstrating a rescue of HD-relevant phenotypes. This is a proof-of-principle study, and further investigations will be required to establish the suitability for therapeutic application.

#### Discussion

We have identified mHTT–LC3 linker compounds that are able to reduce mHTT levels at nanomolar concentrations in HD cells and at  $0.5\,\mathrm{mg\,kg^{-1}}$  by intraperitoneal injection in vivo (Extended Data Table 1). The compounds did not influence wtHTT, which has essential functions—especially during development and young adulthood 36. These features of the compounds are highly desirable for the treatment of HD and potentially for the treatment of other polyQ diseases (Extended Data Fig. 9a–e), although preclinical studies of longitudinal efficacy and safety will be necessary for therapeutic development.

From a broader perspective, we have demonstrated the concept of using small-molecule compounds to target proteins (for example, mHTT) for autophagic degradation by linking them to LC3 (Fig. 5g). We selected mHTT as the target protein because wtHTT provides a good internal control for screening. We identified compounds that interact with both LC3B and mHTT; however, if no such compounds had been

identified, linker compounds could still be generated by conjugating a mHTT-interacting compound and an LC3-interacting compound using the nucleophile-isocyanate reaction used to create the SMMs. The critical next step in developing this concept will be to resolve the core chemical moiety that interacts with LC3 without influencing its function. Comprehensive medicinal chemistry and structural studies are needed to resolve the compound-LC3 interaction interface, which could then be developed to create a general degradation-targeting tool for conjugation with other compounds that interact with specific proteins of interest.

In summary, we have identified mHTT-LC3 linker compounds that are capable of lowering mHTT levels in vivo in an allele-selective manner and demonstrated the possibility of targeting proteins for degradation using autophagosome-tethering compounds, providing new entry points for drug discovery.

#### Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, Supplementary Information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1722-1.

- Kabeya, Y. et al. LC3, a mammalian homologue of yeast Apg8p, is localized in autophagosome membranes after processing. EMBO J. 19, 5720-5728 (2000).
- 2. Scherzinger, F. et al. Huntingtin-encoded polyglutamine expansions form amyloid-like protein aggregates in vitro and in vivo. Cell 90, 549-558 (1997).
- Warrick, J. M. et al. Expanded polyglutamine protein forms nuclear inclusions and causes neural degeneration in Drosophila, Cell 93, 939-949 (1998)
- Fire, A. et al. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans, Nature 391, 806-811 (1998).
- Mali, P. et al. RNA-guided human genome engineering via Cas9. Science 339, 823-826 5. (2013)
- 6. Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. Science 339, 819-823 (2013).
- Winter, G. E. et al. Phthalimide conjugation as a strategy for in vivo target protein 7 degradation. Science 348, 1376-1381 (2015).
- Lu, K., den Brave, F. & Jentsch, S. Pathway choice between proteasomal and autophagic degradation. Autophagy 13, 1799-1800 (2017).
- Mizushima, N., Levine, B., Cuervo, A. M. & Klionsky, D. J. Autophagy fights disease through cellular self-digestion. Nature 451, 1069-1075 (2008).
- Zhu, C. et al. Developing an efficient and general strategy for immobilization of small molecules onto microarrays using isocyanate chemistry. Sensors 16, E378 (2016).
- Fei, Y. et al. Screening small-molecule compound microarrays for protein ligands without fluorescence labeling with a high-throughput scanning microscope. J. Biomed. Opt. 15, 016018 (2010).
- Mangiarini, L. et al. Exon 1 of the HD gene with an expanded CAG repeat is sufficient to cause a progressive neurological phenotype in transgenic mice. Cell 87, 493-506 (1996).
- Liu, H. et al. Nuclear cGAS suppresses DNA repair and promotes tumorigenesis. Nature **563**, 131-136 (2018).
- Landry, J. P. et al. Discovering small molecule ligands of vascular endothelial growth factor that block VEGF-KDR binding using label-free microarray-based assays. Assay Drug Dev. Technol. 11, 326-332 (2013).

- Fei. Y. et al. Characterization of receptor binding profiles of influenza A viruses using an ellipsometry-based label-free glycan microarray assay platform. Biomolecules 5, 1480-1498 (2015).
- Zhu, X. et al. Oblique-incidence reflectivity difference microscope for label-free highthroughput detection of biochemical reactions in a microarray format. Appl. Opt. 46, 1890-1895 (2007)
- Landry, J. P., Zhu, X. D. & Gregg, J. P. Label-free detection of microarrays of biomolecules by oblique-incidence reflectivity difference microscopy. Opt. Lett. 29, 581-583 (2004).
- Zhu, C. et al. Fast focal point correction in prism-coupled total internal reflection scanning imager using an electronically tunable lens. Sensors 18, E524 (2018).
- Menalled, L. B., Sison, J. D., Dragatsis, I., Zeitlin, S. & Chesselet, M. F. Time course of early motor and neuropathological anomalies in a knock-in mouse model of Huntington's disease with 140 CAG repeats. J. Comp. Neurol. 465, 11-26 (2003).
- Bondeson, D. P. et al. Catalytic in vivo protein knockdown by small-molecule PROTACs. Nat. Chem. Biol. 11, 611-617 (2015).
- 21. Miller, J. et al. Identifying polyglutamine protein species in situ that best predict neurodegeneration, Nat. Chem. Biol. 7, 925-934 (2011).
- Fu. Y. et al. A toxic mutant huntingtin species is resistant to selective autophagy. Nat. Chem. Biol. 13, 1152-1154 (2017).
- 23 Baldo B et al. TR-FRFT-based duplex immunoassay reveals an inverse correlation of soluble and aggregated mutant huntingtin in Huntington's disease. Chem. Biol. 19, 264-275 (2012).
- Weiss, A. et al. Single-step detection of mutant huntingtin in animal and human tissues: a bioassay for Huntington's disease. Anal. Biochem. 395, 8-15 (2009).
- Lu, B. et al. Identification of NUB1 as a suppressor of mutant Huntington toxicity via enhanced protein clearance. Nat. Neurosci. 16, 562-570 (2013).
- Mizushima, N. et al. Dissection of autophagosome formation using Apg5-deficient mouse embryonic stem cells. J. Cell Biol. 152, 657-668 (2001).
- Lackey, K. et al. The discovery of potent cRaf1 kinase inhibitors. Bioorg. Med. Chem. Lett. 10, 223-226 (2000).
- Reddy, K. & D'Orazio, A. Highlights from the international conference on molecular targets and cancer therapeutics: discovery, biology, and clinical applications, Philadelphia, PA. ECCO 13-The European Cancer Conference, Paris, France, October 30-November 3, 2005. Clin. Genitourin. Cancer 4, 156-159 (2005).
- Johnson, G. L. & Lapadat, R. Mitogen-activated protein kinase pathways mediated by ERK, JNK, and p38 protein kinases. Science 298, 1911-1912 (2002).
- Tao, W. et al. An inhibitor of the kinesin spindle protein activates the intrinsic apoptotic pathway independently of p53 and de novo protein synthesis. Mol. Cell. Biol. 27, 689-698 (2007).
- Trettel, F. et al. Dominant phenotypes produced by the HD mutation in STHdh<sup>Qm</sup> striatal cells, Hum, Mol. Genet. 9, 2799-2809 (2000).
- Kimura, S., Noda, T. & Yoshimori, T. Dissection of the autophagosome maturation process by a novel reporter protein, tandem fluorescent-tagged LC3. Autophagy 3, 452-460 (2007).
- Zhang, J., Wang, J., Ng, S., Lin, Q. & Shen, H. M. Development of a novel method for quantification of autophagic protein degradation by AHA labeling. Autophagy 10, 901-912 (2014).
- Ni, H. M. et al. Dissecting the dynamic turnover of GFP-LC3 in the autolysosome. Autophagy 7, 188-204 (2011).
- Feng, X., Luo, S. & Lu, B. Conformation polymorphism of polyglutamine proteins. Trends Biochem. Sci. 43, 424-435 (2018).
- Wang, G., Liu, X., Gaertig, M. A., Li, S. & Li, X. J. Ablation of huntingtin in adult neurons is nondeleterious but its depletion in young mice causes acute pancreatitis. Proc. Natl Acad. Sci. USA 113, 3359-3364 (2016).
- Vijayvargia, R. et al. Huntingtin's spherical solenoid structure enables polyglutamine tract-dependent modulation of its structure and function. eLife 5, e11184 (2016).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations

© The Author(s), under exclusive licence to Springer Nature Limited 2019

#### Methods

#### Additional details from figure legends

In Fig. 5a, loss of processes and shrinkage of neurons were observed in HD neurons after BDNF removal. Bar plots show quantification of the area showing TUBB3 signal (TUBB3 area) normalized to the nuclei counts based on DAPI staining. The lower TUBB3 area per cell reflects neuronal processes shrinkage and loss. Data were normalized to the average of wild-type controls. The data were analysed by one-way ANOVA (F(5,60) = 94.78) with post hoc Dunnett's tests. \*\*\*\*P < 0.0001; n indicates the number of independently plated wells.

Figure 5b, the images were captured every 3 h inside the incubator using Incucyte, and the caspase-3 active cells were quantified by the fluorescent-object count per field. The data were analysed by two-way ANOVA (F(43,516)=12.85) with post hoc Dunnett's tests, comparing to the HD\_DMSO group. \*\*\*\*P<0.0001. The numbers in brackets indicate the number of independently plated wells, with four fields per well imaged and averaged for quantification. Three batches were tested and showed consistent results.

In Fig. 5c, left, *Drosophila* expressed full-length HTT proteins (Q128 or Q16) in the nervous system driven by *elav-GAL4*. Seventy-five flies were tested for each group. The data were analysed by log-rank (Mantel–Cox) test, comparing compound-treated groups with DMSO controls in Q128 flies. \*\*\*\*P<0.0001. Figure 5c, right, similar to Fig. 5c, left, but plotting the climbing performance as a function of age after eclosion. Data were analysed by two-way ANOVA (F(4, 275) = 122.1) with post hoc Dunnett's tests, comparing the compound treated groups with the DMSO controls in Q128 flies. Numbers in brackets indicate the number of vials (each containing 15 flies) tested. \*\*\*\*P<0.0001.

In Fig. 5d–f, the numbers in brackets indicate the number of mice tested. The data were analysed by two-way ANOVA with post hoc Dunnett's tests, and P values were calculated for the comparison with the DMSO control. \*\*\*\*P<0.0001. For HD mice, F(2, 195) = 4.963 in rotarod tests, F(2, 195) = 37.31 in balance beam tests, and F(2, 156) = 7.068 in gripping force tests. No significant difference was detected among wild-type mice injected with the different compounds. Investigators were blinded to the compounds and genotypes when performing the experiments. In all panels, graphical data are presented as mean and s.e.m.

#### Compound stamping on the microarray

SMMs containing 3.375 bioactive compounds were used for highthroughput screening of target proteins. The compound library containing 1,527 drugs approved by Food and Drug Administration (FDA) of United States, 1,053 natural products from traditional Chinese medicine, and 795 known inhibitors were stamped onto the SMMs. Each compound was dissolved in DMSO at a concentration of 10 mM and printed in duplicates along vertical direction on homemade phenyl-isocyanate functionalized glass slides with a contact microarray printer (SmartArrayer 136, CapitalBio Corporation). Biotin-BSA at a concentration of 7,600 nM in 1× phosphate-buffered saline (PBS) and biotin-(PEG)<sub>2</sub>-NH<sub>2</sub> at a concentration of 5 mM in DMSO were printed as the inner and outer borders of SMMs, respectively. The diameter of each spot was about 150 μm and spacing between two adjacent spots was 250 µm. The printed SMMs were then dried at 45 °C for 24 h to facilitate covalent bonding of nucleophilic groups of small molecules to isocyanate groups of the functionalized slides. Afterwards, the SMMs were stored in a -20 °C freezer.

#### Expression and purification of recombinant proteins

The human microtubule-associated protein 1 light chain 3- $\beta$  (MAPILC3B (LC3B)) gene (GenBank: NM\_022818.4) was amplified by PCR and cloned into a pGEX-6P1 (GE Healthcare) derived vector pGHT, which is a prokaryotic expression vector reconstructed by adding a His<sub>8</sub> tag and a TEV protease cleavage site before the pGEX-6P1 multiple cloning site. After

sequencing verification, the expression plasmid pGHT-LC3B was introduced into Escherichia coli BL21 (DE3) pLsvS. in which the recombinant GST-LC3B protein was expressed by induction with IPTG. When the bacterial culture reached  $OD_{600} = 0.8$ , its temperature was decreased to 18 °C, and 0.2 mM IPTG was added into the culture for an additional 20 hincubation. The cells were then harvested by centrifugation (6,000g, 4 °C, 15 min) and the cell pellet was suspended in 50 mM Tris-HCl buffer, pH 7.5, with 150 mM NaCl and 5% glycerol. Cells were then disrupted by sonication, followed by centrifugation (20,000g, 4 °C, 60 min). The supernatants were then loaded onto a HisTrap HP column (GE Healthcare, cat. no. 17524701), and eluted with 50 mM Tris-HCl buffer, pH 7.5, containing 150 mM NaCl, 5% glycerol and 300 mM imidazole. The LC3B eluate was then mixed with TEV protease (Sigma, cat. no. T4455: eluted protein: TEV protease = 100:1) and dialysed against the dialysate buffer (50 mM Tris-HCl buffer, pH 7.5, containing 100 mM NaCl) in 4 °C overnight. After TEV protease treatment, the samples were then loaded onto a HisTrap HP column again, the flow through fraction which mainly contains tag removed recombinant LC3B. Afterwards, the proteins were concentrated and further purified by Superose 6 Increase 10/300 GL (GE Healthcare) size-exclusion chromatography. Finally, the purified proteins were concentrated to approximately 10 mg ml<sup>-1</sup> in 50 mM HEPES buffer with 100 mM NaCl for further analysis. The MBP-His<sub>8</sub> and Rpn10 proteins were purified similarly.

The full-length HTT proteins, HTTexon1-MBP, polyQ-sfGFP and sfGFP were purified from mammalian cells. For full-length HTT proteins, the human HTT gene (GenBank: NM 002111.8) with (CAG)<sub>23</sub> or (CAG)<sub>73</sub> (23Q or 73Q for proteins) were de novo synthesized (by Genewiz), sequence validated and then cloned into a modified pCAG vector with an N-terminal protein A tag. The plasmid was transfected to human embryonic kidney E293 cells using polyethylenimine (PEI, from Polysciences, cat. no. 23966). After culture at 37 °C for 48 to 60 h, cells were collected and lysed at 4 °C for 1 h in lysis buffer containing 50 mM Tris-HCl, pH 8.0, 150 mM NaCl, 5% glycerol, 0.5% CHAPS, 3 mM DTT, 1% PMSF, 1 µg ml<sup>-1</sup> pepstatin, 1 µg ml<sup>-1</sup> leupeptin and 1 µg ml<sup>-1</sup> aprotinin, 5 mM ATP and 5 mM MgCl<sub>2</sub>. After centrifugation at 15,000 r.p.m. for 40 min, the supernatants were then incubated with IgG monoclonal antibody-agarose (Smart-lifesciences, cat. no. SA030010) for 2 h and unbound proteins were extensively washed away. The HTT proteins were then digested using TEV protease overnight to remove the protein A tag and eluted protein was further purified by ion exchange and gel filtration chromatography using Mono O and Superose 6 (5/150 GL) columns from GE healthcare. The peak fractions were pooled for further biochemical analysis. The HTTexon1 with 25Q or 72Q cDNA were also de novo synthesized and cloned into a mammalian expression vector pTT5SH8Q2 for large scale production in HEK293T cells. In order to improve the production yield and increase the solubility, a C-terminal MBP tag was added after the HTTexon1 sequences to generate the pTT-HTTexon125Q-MBP and pTT-HTTexon125Q-MBP plasmids. For protein production and purification, the HEK293T cells were transfected by pTT-HTT25QExon1-MBP and pTT-HTT72QExon1-MBP plasmids with linear PEI (PolySciences cat. no. 24765), and then collected after 48 h. The cells were then lysed by sonication in buffer containing 50 mM Tris-HCl, pH7.5, 150 mM NaCl, 20 mM imidazole, 5% glycerol, protease inhibitor cocktail (Sigma) and 50 U ml<sup>-1</sup>benzonase (Sigma). After centrifugation, the supernatants were loaded onto HisTrap HP column (GE Healthcare), and eluted with the buffer containing 50 mM Tris-HCl, pH 7.5, 150 mM NaCl, 300 mM imidazole, 5% glycerol and protease inhibitor cocktail. The MBP tag was not cleaved to avoid precipitation. Afterwards, the proteins were concentrated and further purified by Superose 6 Increase 10/300 GL (GE Healthcare) size-exclusion chromatography.

#### Verifications of the recombinant proteins by MALDI-TOF

The purified LC3B, HTTexon1Q25–MBP, and HTTexon1Q72–MBP proteins were dialysed into 5 mMNH₄Ac by Superose 6 Increase size-exclusion chromatography for linear mode matrix assisted laser desorption

ionization-time of flight mass spectrometry (MALDI–TOF) analysis on a Bruker FLEX MALDI–TOF instrument. A total of 1,500–2,500 scans were averaged for each spectrum using an accelerating voltage of 25 kV. Sinapinic acid (SA, Bruker, cat. no. 820135) was used as the matrices for protein and peptide analyses. Sinapinic acid was made into 20 mg ml $^{-1}$  solutions in 70% acetonitrile, 0.1% trifluoroacetic acid. For the acquisition of spectra from 10,000 to 100,000 amu, 2  $\mu$ l of sample was mixed with 2  $\mu$ l of sinapinic acid solution in an Eppendorf tube, and 2  $\mu$ l of the mixture was loaded onto the MALDI plate. The calibration peptides for this range were BSA (M + 66,431) (Sigma, cat. no. A1933). All spectra were obtained in positive linear mode. The amount of full-length HTT proteins were limited, and thus not validated by MALDI–TOF. Instead, they were further purified by ion exchange and gel filtration chromatography, and validated by Coomassie blue staining (Extended Data Fig. 1e) and western blot (Fig. 4b).

# Verifications of the recombinant LC3B by X-ray diffraction crystallography

Because the deletion of G120 (lipidation site) stabilizes LC3B protein, we used LC3B( $\Delta$ G120) protein to obtain high-resolution diffraction data. Purified LC3B( $\Delta$ G120) protein was concentrated in 20 mM HEPES pH 7.5, 150 mM NaCl. The LC3B( $\Delta$ G120) crystal was grown in reservoir solutions consisting of 0.16 M ammonium sulfate, 0.08 M sodium acetate pH 4.6, 20% (w/v) PEG4000, 20% (v/v) glycerol and 0.01 M taurine.

#### Refinement

The X-ray diffraction data were collected at 100 K in the beamline BL17U1 and BL19U1, SSRF. The wavelength for data collection was 0.97892 Å. Diffraction images were indexed and processed by HKL2000. The structure of LC3B( $\Delta$ G120) (PDB ID: 6J04, 1.90Å) was solved by molecular replacement with the Phaser 2.8 program from the CCP4 crystallography package using PDB structure 1UGM as the search model. The refinement was performed by Refmac 5.5 and Phenix 1.14. There are no Ramachandran outliers to report. The related figure was drawn using PyMOL 2.2.

#### Compound-protein interaction measurements by OI-RD

For high-throughput preliminary screening of target proteins, a SMM was assembled into a fluidic cartridge and washed in situ with a flow of 1× PBS to remove excess unbound small molecules. After washing, the SMM was scanned with a label-free OI-RD scanning microscope to image small molecules immobilized on glass slides. After it was blocked with 7,600 nM BSA in 1× PBS for 30 min, SMM was incubated with the target protein for 2 h. HTTexon1(O25)-MBP at a concentration of 454 nM, HTTexon1(Q72)-MBP at a concentration of 238 nM, and LC3B at a concentration of 680 nM were screened on separate fresh SMMs. OI-RD images were scanned for each operation, including washing, blocking and incubation. The OI-RD difference images (images after incubation – images before incubation) were used for analysis, and vertical bright doublet spots indicated compounds that bind with target proteins in both replicates. Compounds 8F20 and 10O5 were identified to bind to HTTexon1(Q72)-MBP and LC3B, but not to HTTexon1(Q25)-MBP. The binding was further confirmed by kinetics measurements.

To measure binding kinetics of target proteins with compounds, we prepared new SMMs consisting of 8F20, 10O5 and AN2. Six identical microarrays were printed on one glass slide and each compound was printed in triplicates in a single microarray. The printed small SMMs were assembled into a fluidic cartridge with each microarray housed in a separate chamber. Before the binding reaction, the slide was washed insitu with a flow of  $1\times$  PBS to remove excess unbound samples, followed by blocking with 7,600 nM BSA in  $1\times$  PBS for 30 min. For binding kinetics measurement,  $1\times$  PBS was first flowed through a reaction chamber at a flow rate of 0.01 ml min $^{-1}$  for 5 min to acquire the baseline.  $1\times$  PBS was then quickly replaced with the probe solution of the target protein at a flow rate of 2 ml min $^{-1}$  for 9 s followed by a reduced flow rate at 0.01

ml min $^{-1}$ to have the microarray incubated in the probe solution under the flow condition for 35 min (association phase of the reaction). The probe solution was then quickly replaced with 1× PBS at a flow rate of 2 ml min $^{-1}$  for 9 s followed by a reduced flow rate of 0.01 ml min $^{-1}$  to allow dissociation of probe for 30 min (dissociation phase of the reaction). By repeating the binding reactions of the target protein at three different concentrations on separate fresh microarrays, binding curves of compounds with the target protein at three concentrations were recorded with scanning OI-RD microscope. Reaction kinetic rate constants were extracted by fitting the binding curves globally using 1-to-1 Langmuir reaction mode.

#### Compound-protein interaction measurements by MST

The purified recombinant proteins were dialysed into  $1\times$  PBS, and then labelled according to the protocol of Protein labelling kit RED-NHS (Nanotemper, cat. no. L001). All the tested stock compounds ( $10\,\text{mM}$ ) dissolved in DMSO were also diluted into the same buffer for the final MST assay. The MST experiment was performed using Monolith NT.115 instrument (NanoTemper Technologies). Labelled proteins ( $500\,\text{nM}$ ) were mixed with the indicated concentrations of candidate compounds in reaction buffer containing  $20\,\text{mM}$  HEPES, pH 7.4, 150 mM NaCl. The MST data were then collected under 40% infrared laser power and 20% light-emitting diode power. The data were analysed by Nanotemper analysis software (v.1.5.41) and the  $K_{\rm d}$  was determined.

#### cDNA plasmids for transfection in mammalian cells

The pEX-GFP-hLC3WT plasmid was obtained from Addgene (24987) to express LC3B. The pTT-HTTexon1-Q72-MBP-His and pTT-HTTexon1-Q25-MBP-His were generated by subcloning HTTexon1 cDNAs into the mammalian expression vector pTT-MBP-His and then transiently transfected into HeLa cells to express HTTexon1 proteins for the colocalization experiments. The polyQ-GFP sequences (expressing Met-polyQ-sfGFP) were de novo synthesized and subcloned into the pcDNA vector. All plasmids were sequence validated. For transient transfections, the cells were plated at 50% confluence. After 24 h, the cDNAs were transfected with Lipofectamine 2000 (Thermo Fisher Scientific, cat. no. 11668019) using the forward transfection protocol provided by the manufacturer.

#### Cell culture

For mouse primary cortical neuron cultures, cortices were isolated from postnatal day 0 pups following genotyping. Cortices were dissected into cold  $\text{Ca}^{2^+}$  and  $\text{Mg}^{2^+}$ -free PBS buffer. Chopped small pieces were digested in solution containing 2.5% trypsin (Sigma, cat. no. P1005) and DNase I (0.1 mg ml $^{-1}$ , Sigma, cat. no. D5025), for 20–30 min at 37 °C. Tissues were transferred to 10% FBS containing DMEM (Thermo Fisher Scientific, cat. no. 11965) to cease digestion. Neurons were then dissociated by trituration with fire-polished glass pipettes, collected by spinning and plated onto polylysine-coated dishes at  $4\times10^5$  cells per 35-mm dish. The growth medium was composed of Neurobasal A medium (Thermo Fisher Scientific, cat. no. 10888022) with  $1\times$  B-27 (Thermo Fisher Scientific, cat. no. 17504044) and  $1\times$  N2 supplement (Thermo Fisher Scientific, cat. no. 17504048). Cytosine-arabinofuranoside (Sigma, cat. no. C1768) was added at 6  $\mu$ M to inhibit glial growth.

Some of the primary patient fibroblasts were obtained from HD patients (Q47, Q49, Q55) and healthy sibling (WT, Q19) controls in a family with HD from Mongolia. The HD Q68 fibroblast line was obtained from Coriell Cell Repositories. The PD line was obtained from an idiopathic Parkinson's disease patient, and the SCA3 line was obtained from a patient with SCA3 harbouring the ATXN3 expansion mutation (Q74). The studies were approved by The Ethic Community of Institutes of Biomedical Sciences at Fudan University (#28) for obtaining the HD and wild-type patient fibroblasts, and by Huashan Hospital Institutional Review Board at Fudan University (#174) for obtaining the PD and fibroblasts from patients with SCA3. Verbal and written consent was

obtained from patients. The procedures were in compliance with all relevant ethical regulations. The immortalized fibroblasts were generated by infection of lentivirus expressing SV40T. For generation of iPS cells, the primary fibroblasts were transduced with the retroviral STEMCCA polycistronic reprogramming system (Millipore, cat. no. SCR548). The iPS cells were confirmed positive for Tra-1-81, Tra-1-60, SSEA-4 and Nanog by immunofluorescence and flow cytometry. All four vectorencoded transgenes were found to be silenced and the karyotype was normal. iPS cells were cultured in E8 medium (Thermo Fisher Scientific, cat. no. A1517001) on Matrigel (Corning, cat. no. 354277) surface. iPS cells were differentiated to Pax6-expressing primitive neuroepithelia (NE) for 10–12 days in a neural induction medium. Sonic hedgehog (SHH, 200 ng ml<sup>-1</sup>) was added on days 10-25 to induce ventral progenitors. For neuronal differentiation, neural progenitor clusters were dissociated and placed onto poly-ornithine/laminin-coated coverslips at day 26 in Neurobasal medium (Thermo Fisher Scientific, cat. no. 21103049), with 1×B-27 (Thermo Fisher Scientific, cat. no. 17504044), 1× N-2 (Thermo Fisher Scientific, cat. no. 17504048), brain derived neurotrophic factor (BDNF, 20 ng/ml, Protech, cat. no. 450-02), glialderived neurotrophic factor (GDNF, 10 ng/ml, Protech, cat. no. 450-10), insulin-like growth factor 1 (IGF1, 10 ng/ml, Protech, cat. no. 100-11) and Vitamin C (Sigma cat. no. D-0260, 200 ng/ml). The mouse striatal cells (STHdh) were obtained from Coriell Cell Repositories. The HEK293T cells and the HeLa cells were originally obtained from American Type Culture Collection (ATCC). STHdh, HeLa and HEK293T cells were cultured in DMEM (Thermo Fisher Scientific, cat. no. 11965) with 10% (vol/ vol) FBS (Thermo Fisher Scientific, cat. no. 10082-147). Atg5 WT and KO MEFs were from N. Mizushima. All the mammalian cell lines were maintained at 37 °C incubator with 5% CO<sub>2</sub>, except STHdh cells, which were maintained at 33 °C with 5% CO<sub>2</sub>. The cells were tested every two months by a TransDetect PCR Mycroplasma Detection Kit (Transgen Biotech, cat. no. FM311-01) to ensure that they are mycoplasma free. The CellTiter-glo assay was performed to measure cell viability with the indicated compound treatment (Extended Data Fig. 2e) following the protocol provided in the kit (Promega, cat. no. G7570).

#### HD Drosophila models

The nervous system driver line *elav-GAL4* (*c155*), and the HTT-expressing lines *UAS-flHTT-Q16* and the *UAS-flHTT-Q128* (expressing human full-length HTT with 16Q and 128Q, respectively, when crossed to the GAL4 line) lines were obtained from the Bloomington *Drosophila* Stock Center at University of Indiana (http://flystocks.bio.indiana.edu/), and maintained in a 25 °C incubator. Crosses were set up between virgin female flies carrying *elav-GAL4* driver and the *UAS-flHTT-Q16* or *UAS-flHTT-Q128* male flies to generate the desired genotypes.

## **HD** mouse models

The generation and characterization of the Hdh140Q knock-in mice have been previously described  $^{19}$ . Mice were group-housed (up to 5 adult mice per cage) in individually vented cages with a 12 h light/dark cycle. The mouse experiments were carried out following the ARRIVE (Animal Research: Reporting of In vivo Experiments) guidelines, and they were in compliance with all relevant ethical regulations. The Animal Care and Use Committee of the School of Medicine at Fudan University approved the protocol used in animal experiments (Approval 20140904 and 20170223-005).

#### Compound treatment in cells and animals

The compounds used in this study were all commercially available, and quality controlled by the vendors using NMR. 10O5: GW5074 (DC Chemicals; cat.no. DC8810); 8F20: ispinesib (Selleck; cat.no. S1452); ANI: 5-bromo-3-[(4-hydroxyphenyl)methylidene]-2,3-dihydro-1*H*-indol-2-one (Specs; cat. no. AN-655/15003575); AN2: 5,7-dihydroxy-4-phenylcoumarin (ChemDiv; cat.no. D715-2435); GSK923295 (Selleck, cat. no. S7090), BAY1217389 (Selleck, cat.no. S8215), PLX-4720 (Selleck,

cat.no. S1152), Dabrafenib (Selleck, cat.no. S2807), Sorafenib Tosylate (Selleck, cat.no. S1040), rapamycin (Sigma-Aldrich, cat. no. R8781).

For compound treatment in the cells, the compounds were diluted in culture medium to  $10 \times$  concentrations and added to the plated cells: for primary cultured neurons and iPS-cell-derived neurons, the compounds were added 5 days after plating; for patient fibroblasts and other cell lines, the compounds were added 1 day after plating. The cells were then collected 2 days later for measurement of HTT levels. For detection of HTT-LC3 colocalization, the cells were fixed 4 h after compound treatment. For caspase-3 activation detection, the cells were stressed (BDNF removal for iPS-cell-derived neurons) 1 day after compound treatment, and tested at the indicated time points.

For compound treatment in the <code>Drosophila</code>, flies were maintained in standard maize food at 25 °C. For drug feeding, maize media was heated to 45 °C until liquid and distributed into vials. Compounds were freshly prepared in DMSO and added to the media. New adult flies were transferred to vials with 400  $\mu L$  the control (DMSO) or compound-containing food, which was changed every other day.

For compound treatment in mice using intracerebroventricular (icv) injection, the 3-month-old mice were anesthetized using a small animal anaesthesia machine (MSS-3, MSS International) by isoflurane (1.5% solution). We surgically implanted each mouse with a guide cannula directed towards the lateral ventricle. The coordinates for implantation were determined using "The Mouse Brain in Stereotaxic Coordinates" and the guide cannulas were placed at 0.6 mm posterior, 1.5 mm lateral (left), and 1.7 mm dorsal with respect to bregma. A cap with stylus was then inserted into the guide cannula to seal its opening. Mice were then allowed to recover from surgery for a week before being treated. For injection, we first inserted an internal injector cannula so that it extended 0.5 mm beyond the tip of the guide cannula to reach the lateral ventricle. We then injected the mice through the internal injector cannula using a 25 µL syringe (Hamilton 1700 Series Microlitre Syringes, Bonaduz, GR, CH) at a flow rate of 0.25 μL/min powered by a syringe pump (KDS Legato 130) to administer 2 µL of compounds-containing artificial cerebrospinal fluid (ACSF:1 mM glucose, 119 mM NaCl, 2.5 mM KCl, 1.3 mM MgSO<sub>4</sub>, 2.5 mM CaCl<sub>2</sub>, 26.2 mM NaHCO<sub>3</sub>, 1 mM NaH<sub>2</sub>PO<sub>4</sub>) at a concentration of 25 μM (containing 0.125% vol/vol DMSO). 2 μL ACSF containing equivalent amount of DMSO (0.125% vol/vol) was used as the control. The injector cannula was left in place for approximately 60 s to allow for diffusion before placing the caps with stylus back in guide cannulas.

For compound treatment in mice using intraperitoneal (ip) injection, each mouse was weighed. The compounds were diluted with 0.9% NaCl intravenous infusion solution to 0.05 µg/µL (containing 0.011 µg/µL DMSO) and injected into each mouse based on the weight of the mouse (500 µg/kg, containing 110 µg/kg DMSO). As controls, equivalent amount of DMSO was diluted and injected in the same way. Injection of 0.9% NaCl intravenous infusion solution alone was also tested and showed no difference (Extended Data Fig. 9f–h). One injection per day was performed for two weeks before subsequent behavioural experiments or tissue extractions.

Note that in some of the experiments (Figs. 4, 5 and Extended Data Fig. 6b–d), 8F20 and/or AN1 were not tested. 8F20 was not tested because it did not have an effect in vivo by icv-injection (Extended Data Fig. 6a). AN1 was not tested because its structure is highly similar as 10O5 while it had a weaker HTT-lowering effect by icv-injection (Extended Data Fig. 6a).

#### Protein extraction from cells and tissues

For protein extraction from cells, the cell pellets were collected and lysed on ice for 30 min in  $1\times$  PBS + 1% Triton X-100 +  $1\times$  complete protease inhibitor (Sigma-Aldrich, cat. no. 11697498001), sonicated for 10 s, and spun at >20,000 g at 4 °C for 15 min. The supernatants were then loaded and transferred onto nitrocellulose membranes for western blotting. For mouse brain tissues, the mouse striata and cortices were dissected

on ice and grinded by a tissue grinder for 5 min at 60 Hz and lysed on ice for 60 min in brain lysis buffer (50 mM Tris, 250 mM NaCl, 5 mM EDTA, 1% Triton X-100 PH7.4) + 1× complete protease inhibitor (Roche, cat. no. 4693159001). The samples were then sonicated for 10 cycles, 15 s on and 20 s off, and then collected for western blot.

For protein extraction from the mouse brain, the brains were collected and the cortices were acutely dissected on ice and homogenized with a tissue grinder for 5 min at 60 Hz and lysed on ice for 60 min in brain lysis buffer (50 mM Tris, 250 mM NaCl, 5 mM EDTA, 1% (vol/vol) Triton X-100, 1× complete protease inhibitor (Roche, cat. no. 4693159001), pH = 7.4). The samples were then sonicated for 10 cycles, 15 s on and 20 s off, and then collected for western blots, HTRF or dot blots.

For mHTT measurements in the HD *Drosophila* model, the fly heads were collected at the age of 7 days and lysed on ice for 30 min in PBS +1% (vol/vol) Triton X-100  $+1\times$  complete protease inhibitor (Roche, cat. no. 4693159001), sonicated for 10 cycles, 15 s on and 20 s off, and then collected for HTRF.

For all the samples, the protein concentrations were measured to correct the loadings. Different protein concentrations or cell numbers per well were tested to ensure that the signals were in the linear range. Background corrections were performed by subtracting the background signals from blank samples.

#### Western blot and filter trap assays

For western blots, the samples were loaded onto the SDS page gel (5-12%) depending on the molecular weight of the protein of interest). The proteins on the gel were then transferred to the nitrocellulose membranes for blocking and antibody detection. The signal was detected with ECL (Bio-Rad, cat. no. 1705061) after 1 h incubation of the membrane with secondary antibody 1:10,000.

The filter trap assay was performed similarly as previously described  $^{23}, 2\,\mu\text{L}$  (10  $\mu\text{g})$  aliquots of each sample were loaded onto nitrocellulose membranes stacked in the Bio-Dot microfiltration apparatus (Bio-Rad). The membrane was blocked for 1 h with 5% milk and incubated overnight with the antibody 4C9 at a concentration of 1.5  $\mu\text{g}/\mu\text{l}$  in 5% milk diluted in PBS + 0.1% Tween-20. The signal was detected with ECL (Bio-Rad, cat. no. 1705061) after 1 h incubation of the membrane with secondary antibody 1:10,000.

## **HTRF** assays

For HTRF, the assays were similar as previously described  $^{25}$ . The cell or tissue lysates were diluted with the original lysis buffer PBS + 1% (vol/vol) Triton X-100 + 1× complete protease inhibitor (Roche), used for lysing the samples, and then detected with indicated antibody pairs diluted in the HTRF assay buffer (50 mM NaH $_2$ PO $_4$ , 400 mM NaF, 0.1% BSA, 0.05% (vol/vol) Tween-20, 1% (vol/vol) Triton X-100, pH 7.4). The donor antibody concentration was 0.023 ng/µL and the acceptor antibody concentration was 1.4 ng/µL, both in HTRF assay buffer. Different antibody pairs were used for different experiments as indicated in the figure legends. For all the samples, the signals were normalized to the total protein concentrations to ensure equal loadings. Different protein concentrations were pre-tested to ensure that the signals were in the linear range. Background corrections were performed by subtracting the background signals from blank samples.

#### In vitro c-Rafkinase assay

In vitro c-Raf kinase assays were carried out with a c-Raf kinase assay kit (BPS Bioscience, cat. no. 79570). The assays were performed in a 96-well plate according to the manufacturing instruction. The samples and non-reactive negative controls were tested in duplicate according to the instruction.

For details,  $25~\mu L$  of the mixture containing  $5\times$  kinase assay buffer  $(6~\mu L)$ , ATP  $(1~\mu L)$ ,  $5\times$  Raf substrate  $(10~\mu L)$  and water  $(8~\mu L)$  was added to a well.  $5~\mu L$  of water solution containing a test compound at a  $10\times$  desired concentration (DMSO was at 10% at the water solution) was added to the

 $25\,\mu L$  of mixture, and  $20\,\mu L$  of  $1^\times$  kinase assay buffer containing  $2\,ng/\mu L$  c-Raf kinase was added to the mixture in a well to initiate the kinase reaction (at this stage compounds were at  $1^\times$  desired concentration, and DMSO was at 1% concentration). For a non-reactive negative control,  $20\,\mu L$  of  $1^\times$  kinase assay buffer containing no c-Raf was added to the mixture instead. The plate was incubated at  $30\,^\circ C$  for  $45\,\text{min}$ . After the 45-min reaction,  $50\,\mu L$  of kinase Kinase-Glo Max reagent (Promega, cat. no. V6071) was added to each well, and the plate was incubated at room temperature for  $15\,\text{min}$ , in the dark. The plate was read with a microplate reader (BMG Labtech) for luminescence reading. The luminescence reading value measures the levels of ATP remaining, which is inversely related to kinase activity. The non-reactive negative control read value, indicating the level of initial added ATP, subtracted the level of ATP remaining (the luminescence reading) for the value of consumed ATP in the reaction that represents a kinase activity.

#### In vitro pull-down assays

We performed in vitro pull-down assays to test the compounds' influence on the HTT-LC3 interactions. The purified HTTexon1 (with the indicated tags), full-length proteins and the control proteins were incubated with amylose resin (New England BioLabs, cat. no. E8021L) at  $4\,^{\circ}\text{C}$  for 30 min. Immobilized amylose resins were then washed three times with HBS (20 mM HEPES pH7.5, 150 mM NaCl, 0.05% Tween-20). The resulting amylose resins containing about 10  $\mu g$  of MBP-fused proteins were incubated with the indicated compounds (1  $\mu M$  for 1005 and 100 nM for AN2) or the DMSO control at the same volume in 300  $\mu l$  of HBS at  $4\,^{\circ}\text{C}$  for  $1\,h$  using sample mixer. 40  $\mu g$  of purified LC3B protein were then added and incubated at  $4\,^{\circ}\text{C}$  for another 2 h using sample mixer. The resin-bound proteins were eluted with  $40\,\mu l$  maltose buffer (10 mM maltose, 20 mM HEPES, 150 mM NaCl, pH7.5) and then added with 20  $\mu l$  SDS-PAGE sample loading buffer. Samples were then analysed by SDS-PAGE and western blots.

GST pulldown was performed as the same procedures described above, except that GST-fused LC3B was immobilized onto magnetic conjugated GST mouse mAb beads (Cell Signaling Technology, cat. no.11847S) and eluted with SDS-PAGE protein loading buffer by vortex according to the instruction manual.

In Fig. 4a, b, for MBP pull-down (Fig. 4a), purified HTTexon1-MBP (10 µg) or MBP (10 µg) bound MBP resin were incubated with the purified LC3B protein (40 µg) and the indicated compounds. The HTTexon1-MBP or the MBP proteins were pulled down and the eluates were tested for co-precipitated LC3B. Four per cent of the total eluate was loaded in each lane, and the input:pull-down loading ratio was 100%. Both 1005 and AN2 enhanced LC3B's interaction with HTTexon1-O72-MBP, but not HTTexon1-Q25-MBP. Note that the MBP blot signals were much weaker for the Q72 protein, possibly because recognition of the MBP tag by the antibody was affected in the fusion protein. Meanwhile, data interpretation was not influenced, because compound treatments did not alter the MBP signals for the Q72 protein (last three lanes). The GST pull-down (Fig. 4b) was performed similarly, except using full-length HTT-Q73 or full-length HTT-Q23 (both without fusion tags) and GST-LC3B proteins for the in vitro GST pull-down experiments to precipitate GST-LC3B or GST alone with its binding proteins, and then eluted for detection. Note that the pull-down is in the reverse direction of the pull-down in (Fig. 4a). The input:pull-down loading ratio for the GST blot was 100%, whereas the ratio for the HTT blot was 10% to avoid overexposure of the input. Both 10O5 and AN2 enhanced LC3B's interaction with the full-length HTT-Q73 but not the full-length HTT-Q23 protein.

#### Imaging-based autophagy assays

Analysis of GFP-LC3 puncta for measuring autophagosomes: HeLa cells stably expressing GFP-LC3 were generated by transfection of pEGFP C1-LC3, and positive clones were selected by  $500 \, \mu g/ml$  G418. The cells were then treated with vehicle (DMSO, 0.1%), 1005, or AN2 for the indicated concentration, chloroquine (CQ,  $20 \, \mu M$ ) treatment was used as

a control. After 24 h, cells were fixed in 4% paraformaldehyde (PFA) for 10 min. Images were acquired with confocal microscopy (Leica SP8) by the observer blinded to the identity of the slides. The number and size of GFP vesicles per cell was determined by Image J software. Images were processed with the despeckle function to decrease the noise, and a threshold was set to highlight puncta. Cells were selected by the freehand drawing tool. The analyse-particle function was used for the sizes and numbers of GFP puncta.

The mRFP-GFP-LC3 assay: this assay allows us to monitor autophagosome synthesis and maturation/fusion by labelling autophagosomes (green and red) and autolysosomes (red), since the low lysosomal pH in autolysosomes quenches the GFP signals  $^{32}$ . HeLa cells stably expressing mRFP-GFP-LC3 were treated with vehicle (DMSO, 0.1%), 10O5, or AN2 for the indicated concentration, bafA1 (10 nM) treatment was used as a control. After 24 h, cells were fixed in 4% PFA for 10 min. Images were acquired with confocal microscopy (Leica SP8) by the observer blinded to the identity of the slides. The green and red single channel images were analysed by Image J to quantify green and red puncta in the same way as in the GFP-LC3 assay described above.

In Fig. 4c, d, representative confocal microscopy images (scale bar, 10 μm) and quantifications of the co-localization between HTTexon1-MBP-His (red, detected by anti-His immunofluorescence) and LC3B-GFP (green, detected by GFP fluorescence directly) in transiently transfected HeLa cells (Fig. 4c) or between endogenous mHTT and LC3-II in the HD-knock-in mouse striatal cells (STHdh  $^{\rm Q111/Q111})$  (Fig. 4d). For overexpressed proteins (Fig. 4c), the cells transfected with LC3B-GFP alone or HTTexon1-MBP-His alone were imaged in both channels to ensure the specificity of the signals (top). The white arrows indicate representative co-localization puncta. Parts of the images have been magnified to show co-localization puncta more clearly (indicated by orange arrows). Since the puncta were obvious, co-localization was analysed by counting the red<sup>+</sup>green<sup>+</sup> (yellow) and the total red<sup>+</sup> puncta directly, and then calculating the ratio for each cell. Blind analysis was performed for quantifications. For endogenous proteins (Fig. 4d), mHTT was detected by the anti-HTT antibody 2166, and the endogenous LC3-II was detected by an anti-LC3 antibody that has been reported to specifically detect LC3-II<sup>32</sup>. Since the signals of endogenous proteins were more dispersed, the co-localization analysis was performed blindly by measuring the red<sup>+</sup>green<sup>+</sup> (yellow) and the total red<sup>+</sup> pixels using Imagel, and then calculate the ratio for each cell.

#### Detection of long-lived proteins by click-chemistry

As an indicator of autophagy activity, the degradation of long-lived proteins was measured similarly as previously reported<sup>33</sup>. In brief, the HeLa cells with 70-80% confluency in a 6-well plate were washed with warm PBS and cultured in Met-free DMEM (Thermo Fisher Scientific, cat. no.21013) added with dialysed FBS for 1 h to deplete intracellular free Met reserves. The Met analogue L-AHA (50 µM) was then added to label the proteins for 18 h. After labelling, the cells were washed with PBS and cultured in regular culture medium containing 10×L-Met (2 mM) for 2 h to chase out short-lived proteins. The cells were then treated with the compounds versus the DMSO controls for 6 h before cell lysis and protein extraction. For the starvation sample, the culture medium was replaced with EBSS (Thermo Fisher Scientific, cat. no. 24010043) for 6 h. The protein lysates were then used for the click reaction by the Click-it reaction kit (Click Chemistry tools, cat. no. C1001) following manufacturer's instructions, and the remaining L-AHA containing long-lived proteins were then conjugated with biotin. These proteins were then analysed by electrophoresis and detected by the HRP-conjugated streptavidin (Beyotime, cat. no. A0303).

#### Immunofluorescence and caspase-3 imaging

For immunofluorescence of cultured cells, cells were fixed in 4% PFA for 10 min after washing with 1× PBS three times, and then washing and permeabilized in 0.5% (vol/vol) TritonX-100 for 10 min. The cells were

then blocked in blocking buffer (4% BSA + 0.1% (vol/vol) Triton X-100 in 1× PBS) for 30 min and incubated overnight at 4 °C with primary antibodies, and then washed three times with blocking buffer and incubated with secondary antibody at room temperature for 1h. Coverslips were then washed three times, stained with 0.5 mg/ml DAPI for 5 min at room temperature, and then mounted in vectashield mounting medium (Vector, cat.no. H-1002). Images were taken by Zeiss Axio Vert A1 confocal microscopes and analysed blindly by Imagel for co-localization and TUBB3 quantifications. For co-localization experiments of transfected HeLa cells (Fig. 4c), the GFP signals were used to detect GFP-LC3B, and anti-His was used to detect HTTexon1-MBP-His proteins. Empty vector transfected cells were imaged to ensure the specificity of the signals. The co-localization was analysed by calculating the ratio between overlapping puncta and the HTT (red) puncta for each cell, and the puncta numbers were counted blindly. For co-localization experiments of STHdh<sup>Q111/Q111</sup> cells, the endogenous mHTT protein was stained with the HTT antibody (Millipore, cat. no. MAB2166), and autophagosomes were stained with the LC3B antibody (Thermo Fisher Scientific, cat. no. 700712), which preferentially detects LC3-II<sup>38</sup>. The co-localization was analysed by ImageJ to calculate the ratio between overlapping pixels and the HTT (red) positive pixels, because the signals of the endogenous proteins were more dispersed and could not be counted accurately. For TUBB3, the total area of TUBB3 signals and the DAPI counts were analysed by ImageJ. The former is then divided by the latter to calculate the averaged area of TUBB3 in each neuron as an index for neurodegeneration in vitro.

For caspase-3 activity measurements of the iPS-cell-derived neurons, the NucView 488 caspase-3 dye (Biotium, cat. no. 30029) was used for the caspase 3 activity detection as an indicator for apoptosis. The images were then taken every 3 h using the Incucyte technology (Essen Bioscience, IncuCyte FLR), which takes images of 4 different fields in each well inside the cell culture incubator. The quantification was performed by the Incucyte 2011A software, which identified the green fluorescent puncta and quantified the fluorescent object count per field. The 4 fields per well were quantified and averaged, and 4 independent wells were used for statistical analysis.

#### **Antibodies**

Antibodies used for western blots, HTRF and/or immunofluorescence/immunohistochemistry are as follows: the HTT antibodies 2B7<sup>24</sup>, ab1<sup>39</sup> and MW1<sup>40</sup> have been described previously: commercially purchased antibodies include HTT antibody 2166 (Millipore, cat. no. MAB2166), anti-polyQ antibody 3B5H10 (Sigma, cat. no. P1874), anti-HTT antibody (D7F7)XP (Cell Signaling Technologies, cat. no. 5656 s), anti-β-tubulin (Abcam, cat. no. ab6046), anti-TUBB3 (Biolegends (previously Covance), cat. no. 801202), anti-ATXN3 (Millipore, cat. no. MAB5360); anti-Gapdh (Proteintech, cat. no. 60004-1), anti-NBR1 (Thermo Fisher Scientific, cat. no. PA5-54660), anti-β-actin (Beyotime, cat. no. AA128); anti-TBP (Abcam, cat. no. ab818); anti-P62 (Thermo Fisher Scientific, cat. no. PA5-27247); anti-spectrin (Millipore, cat. no. MAB1622); anti-Ncoa4 (Santa cruz, cat.no. sc-373739); anti-GST (ProteinTech, cat. no. HRP-66001); anti-GFP (Cell Signaling Technologies, cat. no. 2956); anti-MBP (ProteinTech, cat. no. 15089-1-AP); anti-His (Beyotime, cat. no. AH367); anti-BUBR1 (BD Transduction, cat.no, 612503); anti-phospho-p44/42 MAPK (ERK1/2) and anti-phospho-MEK1/2 in the Phospho-Erk1/2 Pathway Sampler Kit (Cell Signaling Technology, cat.no. 9911); anti-LC3B (Thermo Fisher Scientific, cat. no. PA1-16930 (for western blot) and cat. no. 700712 (for immunofluorescence)). All the antibodies used for immunofluorescence in this study have been validated by knock-down experiments. All the HTT, polyQ and ATXN3 antibodies used for HTRF and/or western blots have been validated by knock-down experiments and by comparing the signals from different genotypes in previous studies from us and others. All the other antibodies have been validated by previous literature or the vendor.

#### Compound detection in vivo in brain tissue from ip-injected mice

The experiments were performed by the SIM-Servier joint laboratory. The mice, ip-injected with DMSO or the indicated compounds, were anesthetized by chloral hydrate (200 uL/kg of 10% stock) at indicated time points, and the heart blood was collected by vacuum blood collection tubes. The heart blood samples were further spun at 10,000 r.p.m. for 5 min to generate the heart plasma. The mice were then perfused with 1× PBS to remove the blood. The mice were then euthanized and the brain samples were dissected. Five times the volume of methanol: acetonitrile (50:50, vol/vol) were added to each sample, which was then homogenized. Following ultrasonic treatment for 15 min, the homogenates were centrifuged for 5 min, then 20 µL supernatant liquid was mixed with 20 uL water for 30 s before injection. Linear range of 1005 was 10-30,000 ng/mL, and the linear range of AN2 was 0.3-10,000 ng/mL. The LC-MS/MS analyses were performed on an Acquity ultra performance liquid chromatography (UPLC) system (Waters Corporation) coupled to a Xevo TQ-S mass spectrometer (Waters Corporation). Chromatographic separation was performed using an Acquity UPLC BEH C18 (1.7  $\mu$ m 2.1  $\times$  50 mm) column supplied by Waters at a flow of 0.5 mL/min. Gradient elution was used with a mobile phase composed of solvent A (water containing 0.1% formic acid and 5 mM NH₄AC) and solvent B (acetonitrile: methanol (9:1,vol/vol) containing 0.1% formic acid).

#### **Proteomics analysis**

Samples were analysed on Orbitrap Fusion Lumos mass spectrometers (Thermo Fisher Scientific) coupled with an Easy-nLC 1000 nanoflow LC system (Thermo Fisher Scientific). Dried peptide samples were re-dissolved in Solvent A (0.1% formic acid in water) and loaded to a trap column (100 μm × 2 cm; particle size, 3 μm; pore size, 120 Å; SunChrom) with a max pressure of 280 bar using Solvent A, then separated on a 150 μm × 15 cm silica microcolumn (particle size, 1.9 μm; pore size, 120 Å; SunChrom) with a gradient of 5–35% mobile phase B (acetonitrile and 0.1% formic acid) at a flow rate of 600 nl min<sup>-1</sup> for 75 min. The FAIMS device was placed before the mass spectrometer. FAIMS separation was performed with the following settings: inner electrode temperature =  $100 \, ^{\circ}$ C, outer electrode temperature =  $100 \, ^{\circ}$ C, carrier gas flow =  $4.6 \, \text{l min}^{-1}$ , dispersion voltage = -5,000 V, entrance plate voltage = 250 V. The FAIMS carrier gas is N<sub>2</sub> only. The noted CVs were applied to the FAIMS electrodes. Each of the selected CVs was applied to sequential survey scans and MS/MS cycles (1s): the MS/MS CV was always paired with the appropriate CV from the corresponding survey scan. For detection with Fusion or Fusion Lumos mass spectrometry, a precursor scan was carried out in the Orbitrap by scanning m/z300–1400 with a resolution of 120,000. The most intense ions selected under top-speed mode were isolated in Quadrupole with a  $1.6 \, m/z$  window and fragmented by higher energy collisional dissociation (HCD) with normalized collision energy of 30%, then measured in the linear ion trap using the rapid ion trap scan rate. Automatic gain control targets were 5 × 10<sup>5</sup> ions with a max injection time of 50 ms for full scans and  $1 \times 10^4$  with 35 ms for MS/MS scans. Dynamic exclusion time was set at 18 s. Data were acquired using the Xcalibur software (Thermo Scientific).

Raw files were searched against the human National Center for Biotechnology Information (NCBI) Refseq protein database (updated on 04-07-2013, 32,015 entries) by Mascot 2.3 (Matrix Science) implemented on Proteome Discoverer 2.2 (Thermo Scientific). The mass tolerances were 20 ppm for precursor and 0.5 Da for product ions for Fusion Lumos. Up to two missed cleavages were allowed. The search engine set cysteine carbamidomethylation as a fixed modification and N-acetylation, oxidation of methionine as variable modifications. Precursor ion score charges were limited to +2, +3, and +4. The data were also searched against a decoy database so that protein identifications were accepted at a false discovery rate of 1%. Label-free protein quantifications were calculated using a label-free, intensity-based absolute quantification (iBAQ) approach.

Proteins with at least 2 unique peptides with 1% FDR at the peptide level and Mascot ion score greater than 20 were selected for further analysis. The file used for protein inference and protein FDR calculation was derived from Mascot search results, and the peptide spectrum match (PSM) was filtered via Percolator and customized parameters, and then the proteins were assembled. The protein FDR was calculated depending on the ratio of NPD (the number of assembled proteins from decoy database searches) and NPT (the number of assembled proteins from target database searches). The FOT was used to represent the normalized abundance of a particular protein across samples. FOT was defined as a protein's iBAQ divided by the total iBAQ of all identified proteins within one sample. The FOT was multiplied by 10<sup>5</sup> for the ease of presentation. Only the proteins detection in all compared samples were used for comparison.

#### Behavioural and lifespan experiments in HD Drosophila models

For behavioural experiments, we placed 15 age-matched virgin female flies in an empty vial and tapped them down. The percentage of flies that climbed past a 7-cm-high line after 15 s was recorded. The mean of five observations is plotted for each vial on each day, and data from multiple vials containing different batches of flies were plotted and analysed by two-way ANOVA tests. The flies were randomly placed into each tube. For lifespan measurements, we placed 75 age-matched virgin female flies in an empty plastic vial and recorded the survival situation for each vial on each day. For both behavioural and lifespan measurement experiments, the person who performed the experiments were blinded to the drugs fed until data analysis.

#### Behavioural experiments in HD mouse models

All the behavioural experiments were performed during the light phase and the experimenters were blinded to the compound treatment and the genotype of each mouse. Both males and females were used. All the mice were kept in the behavioural test room in dim red light for 1h before starting the experiments. For rotarod experiments, mice were pre-trained on 3 consecutive days on the rotarod rotating at 4 r.p.m. for 2 min. Mice were then tested for five days at an accelerating speed ranging from 4 to 40 r.p.m. within 2 min. Each performance was recorded as the time in seconds spent on the rotating rod until falling off or until the end of the task. Each test included three repetitions with an inter-trial interval of 60 min in order to reduce stress and fatigue, and the means from these three runs were analysed for each mouse. The balance beam test was run using a 2-cm-thick metre stick suspended from a platform on both sides by metal grips. The total length is 100 cm. There was a bright light at the starting point and a dark box with food at the endpoint. The total time for each mouse to walk through the beam was recorded. For gripping force measurements, mice were allowed to grip the metal grids of a grip meter (Ametek Chatillon) with their forelimbs, and they were gently pulled backwards by the tail until they could no longer hold the grids. The peak grip strength observed in 10 trials was recorded.

#### **Statistics**

To ensure to reach a statistical power >0.8, power analyses were performed for each assay based on estimated values by PASS 16 (https://www.ncss.com/software/pass/) before experiments. Estimation was based on our previously published results on similar experiments and preliminary experiments. The effect size was also estimated by Cohen's d, two means divided by the standard deviation for the data. The power analysis suggested  $n \ge 3$  for mHTT level measurements and  $n \ge 5$  for behavioural experiments. In all the experiments we performed, we have used a larger n than these numbers in case the effect was smaller than preliminary results, and we also performed post-experiment power analyses to ensure that power  $\ge 0.8$  for all the significant differences. Statistical comparisons between two groups were conducted by the unpaired two-tailed t-tests. Statistical comparisons among multiple

groups were conducted by one-way ANOVA tests and post hoc tests for the indicated comparisons (Dunnett's tests for comparison with a single control, and Bonferroni's tests for comparisons among different groups). Statistical comparisons for series of data collected at different time points were conducted by two-way ANOVA tests. The similarity of variances between groups to be compared was tested when performing statistics in GraphPad Prism 7 and Microsoft Excel 2016. Normality of data sets was assumed for ANOVA and *t*-tests, and was tested by Shapiro–Wilk tests. When the data were significantly different from normal distribution, nonparametric tests were used for statistical analysis. All statistical tests were unpaired and two-tailed.

For the in vivo experiments in the mouse, randomization was performed by assigning random numbers. For the *Drosophila* experiments, the flies were randomly distributed into the vials after anesthesia.

#### **Reporting summary**

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

#### **Data availability**

The protein structure data has been uploaded to the Protein Data Bank with accession number 6JO4. Source data for all figure plots are provided with the paper. The full gel blots and the proteomics data sets have been provided in the Supplementary Information. The data that support the findings of this study are available from the corresponding authors upon reasonable request.

- Hancock, M. K., Hermanson, S. B. & Dolman, N. J. A quantitative TR-FRET plate reader immunoassay for measuring autophagy. Autophagy 8, 1227–1244 (2012).
- Sapp, E. et al. Native mutant huntingtin in human brain: evidence for prevalence of fulllength monomer. J. Biol. Chem. 287, 13487–13499 (2012).

40. Ko, J., Ou, S. & Patterson, P. H. New anti-huntingtin monoclonal antibodies: implications for huntingtin conformation and its binding proteins. *Brain Res. Bull.* **56**, 319–329 (2001).

Acknowledgements We thank J. Lu, M. Jiang, L. Liu and Q. Huang for their technical support with mouse behavioural experiments, Y. Xu for technical support with protein purification and H. Saiyin for help with obtaining human patient fibroblasts. We thank the following for funding support: National Key Research and Development Program of China (2016YFC0905100), National Natural Science Foundation of China (8192500069, 81870990, 31961130379, 91649105, 31470764, 91527305 and 61505032), Science and Technology Commission of Shanghai Municipality (18410722100), Natural Science Foundation of Shanghai (192R1405200), Shanghai Municipal Science and Technology Major Project (No.2018SHZDZX01), ZJLab and Hereditary Disease Foundation.

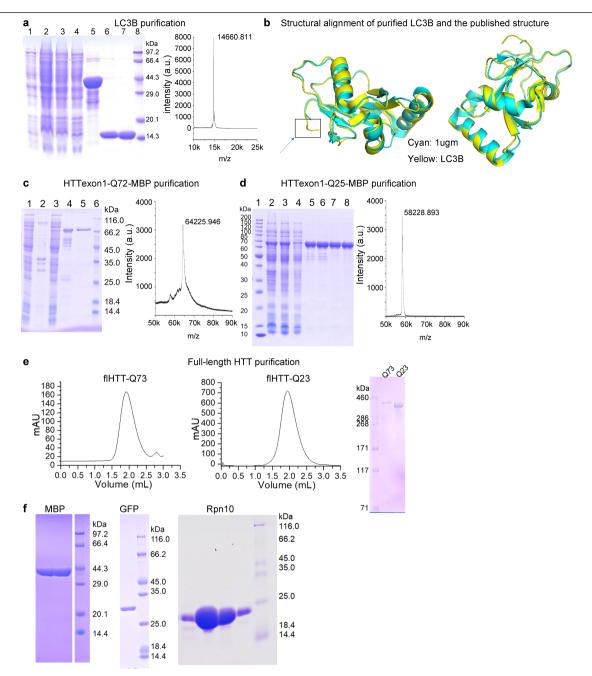
Author contributions B.L. conceived the idea, initiated the project, designed experiments, analysed data and wrote the manuscript. Y.F. and C.Z. performed the OI-RD screening and  $K_{on}/K_{off}$  measurements with data analysis. Y. Ding, Z.W., J.L. and C.G. performed protein purification, in vitro pull-down and structural biology experiments with data analysis. With help from others for blinding, Z.L. performed HTT measurements in cells and in mouse models, the mouse phenotype experiments, the autophagy-related mechanistic experiments, the control protein measurements and neurotoxicity measurements. C.W. replicated the HTT measurement and autophagy-related mechanistic experiments performed by Z.L. and performed additional HTT measurement and phenotypic experiments in HD fly models, patient iPS-cell-derived neurons and MEFs, as well as all measurements of other polyQ proteins and all the MST experiments. Y. Dang provided the compound library for the screen. T.S. and C.D. performed proteomics experiments and analysis. S.L. and Y.Y. performed the measurements of autophagy function. L.M., Y.S. and J.W. provided and characterized the patient cell lines. X.S. did the initial subcloning of full-length HTT and found the explanation for the observed 'hook effects', C.L. performed biostatistical analysis, M.D. and Y.M. helped with designing the experiments and interpreting the data

Competing interests B.L., Y.F., Y. Ding and Y. Dang have filed two patents together on the basis of this study to the State Intellectual Property Office of China (201910180674.7 and 201910180717.1).

#### Additional information

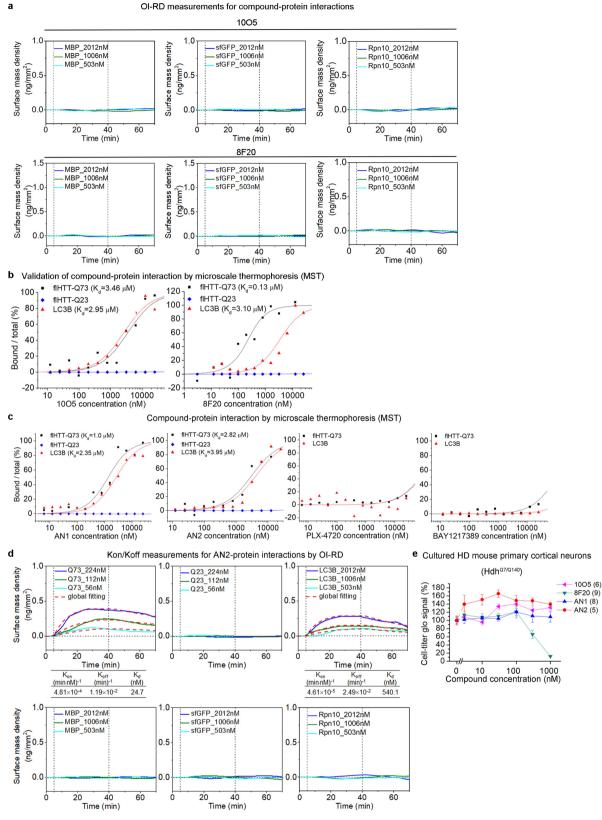
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1792-1

Correspondence and requests for materials should be addressed to Y.D., Y.F. or B.L. Peer review information Nature thanks David Rubinsztein, Huda Yahya Zoghbi and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Reprints and permissions information is available at http://www.nature.com/reprints.



Extended Data Fig. 1| Protein purifications. a, SDS-PAGE and linear mode MALDI-TOF mass spectrometry analysis of the expression and purification of recombinant LC3B protein. Left, SDS-PAGE: lane 1, the whole cell lysate before induction; lane 2, the whole cell lysate after induction; lane 3, the supernatants of induced cells; lane 4, the flow through fraction of Ni-NTA chromatography; lane 5, the eluates of Ni-NTA chromatography (GST-His\_B-LC3); lane 6, LC3B eluate after removal of GST-His\_8 tag by TEV protease; lane 7, the eluates of size-exclusion chromatography; lane 8, molecular weight marker. Right, m/z peak of recombinant LC3B is 14,660.811, consistent with theoretical calculations. b, Structural alignment of purified recombinant LC3B( $\Delta$ G120) (PDB ID: 6J04, yellow) with published LC3B structure (PDB ID: 1UGM, cyan) by PyMOL. c, d, SDS-PAGE and linear mode MALDI-TOF mass spectrometry analysis of the HTTexon1 proteins. c, Left, SDS-PAGE for HTTexon1(Q72)-MBP: lane 1, the supernatants of induced cells; lane 2, the insoluble fraction of induced cells;

lane 3, the flow through fraction of Ni-NTA chromatography; lane 4, the eluates of Ni-NTA chromatography; lane 5, the eluates of size-exclusion chromatography; lane 6, molecular weight marker.  ${\bf d}$ , Left, SDS-PAGE for HTTexon1(Q25)-MBP: lane 1, molecular weight marker; lane 2, the induced cell lysate; lane 3, the supernatant fraction of induced cells; lane 4, the flow-through fraction of Ni-NTA chromatography; lanes 5 and 6, the eluates of Ni-NTA chromatography; lanes 7 and 8, the eluates of size-exclusion chromatography. The m/z peaks of HTTexon1(Q72)-MBP ( ${\bf c}$ , right) and HTTexon1(Q25)-MBP ( ${\bf d}$ , right) are 64,225.946 and 58,228.893, consistent with theoretical calculations.  ${\bf e}$ , Left and middle, size-exclusion chromatography of the recombinant full-length HTT(Q73) (flHTT-Q73) and HTT(Q23) (flHTT-Q23) proteins using Superose 65/150 GL. The major peak fractions were collected pooled together for the SDS-PAGE analysis (right).  ${\bf f}$ , SDS-PAGE analysis of purified MBP-His $_{\bf s}$  (MBP), sfGFP (GFP) and Rpn10 proteins.

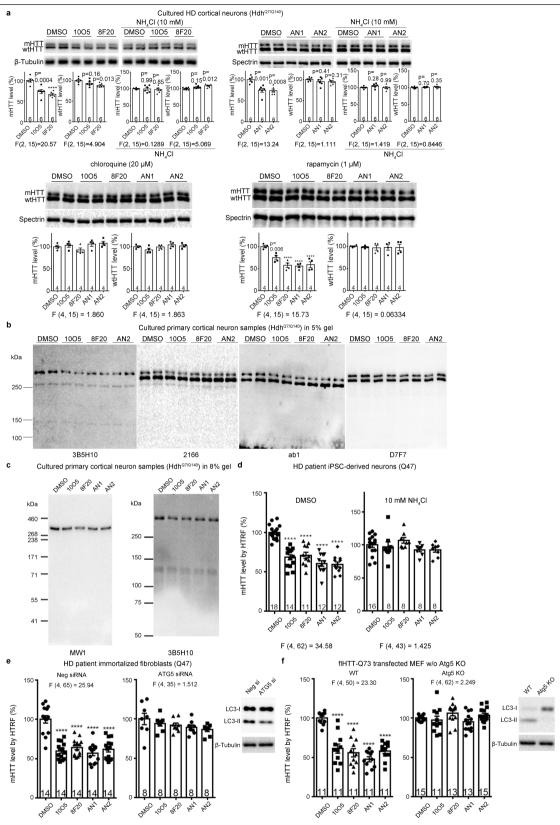


 $\textbf{Extended Data Fig. 2} | See \ next \ page \ for \ caption.$ 

# Extended Data Fig. 2 | Negative controls for OI-RD measurements and validation of the compounds' interaction with HTT and LC3 by MST.

**a**, Similar to Fig. 1c–e, but for negative control proteins MBP–His $_8$  (MBP), sfGFP and Rpn10 (Rpn10). Association–dissociation curves of surface immobilized compounds 8F20 and 10O5 with these proteins were measured by Ol-RD, and no compound–protein interactions were detected. For all association–dissociation curves, vertical dashed lines mark the starts of association and dissociation phases of the binding event. **b**, Binding of 10O5 and 8F20 to full-length HTT(Q73) (flHTT(Q73), black dots) or LC3B (red dots) in standard treated capillaries measured by MST. The compound-bound protein fractions (bound/total) were calculated from the MST signals ( $F_{\text{norm}}$ ) at each compound concentration, as well as the bound ( $F_{\text{norm}}$ , bound, set as 100%) and the unbound ( $F_{\text{norm}}$ \_unbound) set as 0%) MST signals: bound/total = ( $F_{\text{norm}} - F_{\text{norm}}$ \_unbound)/ ( $F_{\text{norm}}$ \_bound– $F_{\text{norm}}$ \_unbound) × 100%. The fitted curves and  $K_{\text{d}}$  values calculated by Nanotemper analysis software (v.1.5.41) for flHTT(Q73) and LC3B are indicated in each panel. Consistent with the Ol-RD measurements (Fig. 1e), no binding was observed for the flHTT(Q23) protein (blue dots). The MST experiments

were repeated more than three times and showed consistent results. c, Similar to **b**, except using the compounds indicated on the x axis. MST measurements of the binding of indicated compounds to full-length HTT(Q73) (flHTT-Q73),  $full-length\,HTT(Q23)\,(flHTT-Q23)\,and\,LC3B\,in\,standard\,treated\,capillaries.\,The$ proteins tested are indicated in the legends. **d**, Similar to Fig. 1c-e, but plotting the association-dissociation curves of surface immobilized compound AN2 with full-length HTT(Q73) (Q73), or full-length HTT(Q23) (Q23), LC3B or the negative-control proteins MBP-His<sub>8</sub> (MBP), sfGFP and Rpn10. For all association-dissociation curves, vertical dashed lines mark the starts of association and dissociation phases of the binding event. The red dashed lines are global fits to a Langmuir reaction model with the global fitting parameters listed at the bottom of each plot. No binding signals were observed for fulllength HTT(O23) proteins, and thus the parameters were not presented, e. Cell viability measurement of cultured HD neurons measured by the CellTiter-glo assay. No toxicity was observed within the concentration range presented in Fig. 2, although the compound 8F20 became toxic to the cells when the concentration reached 300 nM.

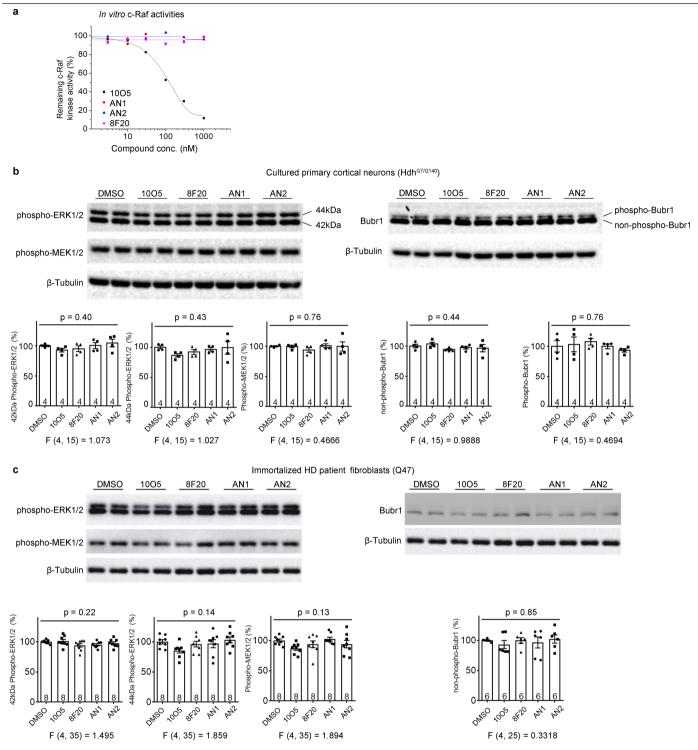


 $\textbf{Extended Data Fig. 3} | See \ next \ page \ for \ caption.$ 

 $Extended \ Data \ Fig. 3 \ | \ mHTT-lowering \ effects \ by \ mHTT-linker \ compounds \\ could \ be \ detected \ by \ multiple \ antibodies \ and \ were \ dependent \ on$ 

autophagy. a, Representative western blots (HTT detected by the 2166 antibody) and quantifications of compound-treated cultured cortical neurons from  $Hdh^{Q7/Q140}$  HD-knock-in mice. The neurons were treated with the indicated compounds (100 nM for 1005, 8F20 and AN1; 50 nM for AN2) with or without the autophagy inhibitor NH4Cl (top) or chloroquine (bottom left), or the autophagy activator rapamycin (bottom right). The same amount of culture medium was added in the controls (top). The statistical analysis was performed by one-way ANOVA with post hoc Dunnett's tests, and the F, degree of freedom and post hoc P values are indicated in each bar plot. **b**, Western blots using indicated HTT or polyQ antibodies for samples from cultured cortical neurons treated with the indicated compounds: 1005 (100 nM), 8F20 (100 nM) or AN2 (50 nM). The HTT gel blots presented in Fig. 2d (right) were cropped from first four blots. The low molecular weight bands were run out in these blots so that the wtHTT and mHTT could be better separated. Note that the weak bands just above 250 kDa in the first two blots were leftover signals from the spectrin blotting. The spectrin signals were too strong to be stripped completely. c, Western blots using the antibody MW1 or 3B5H10, which detects mHTT specifically. We ensured that the relatively low-molecular-weight proteins did not run out of the gels. No increase of potential polyQ-containing mHTT N-terminal fragments was observed. d, iPS-cell-derived striatal neurons from a patient with HD (Q47) were treated with the indicated compounds (100 nM, with 0.1% DMSO) in presence of an additional 0.1% DMSO or 10 mM NH<sub>4</sub>Cl, and

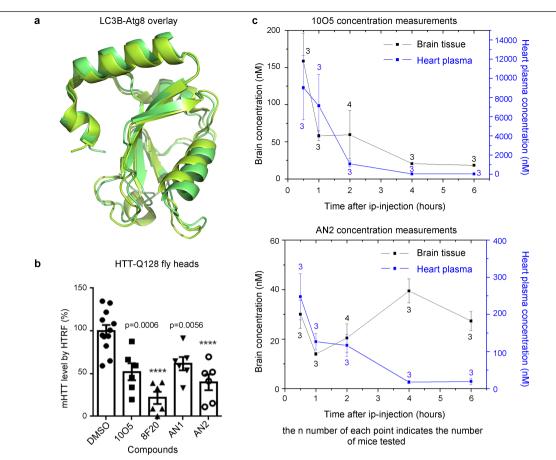
the mHTT levels were measured by HTRF using the 2B7/MW1 antibody pair. All signals were normalized to the averaged signals from the DMSO control group. The statistical analysis was performed by one-way ANOVA with post hoc Dunnett's tests, and F, degree of freedom and post hoc P values are indicated in each bar plot. \*\*\*\*P < 0.0001. The post hoc analysis was not performed if the ANOVA tests did not show significance (P > 0.05). **e**, Immortalized fibroblasts from a patient with HD (Q47) were transfected with the non-targeting control siRNA (Neg siRNA) or the ATG5 siRNA (target sequence, GCCUGUAUGUACUGCUUUA; ATGS mRNA was knocked down to 17.7  $\pm$  3.0%, n=3, as tested by reverse transcription with quantitative PCR), and then treated after 24 h with the indicated compounds (100 nM) for a further 48 h. mHTT levels were then measured by HTRF using the 2B7/MW1 antibody pair. All signals were normalized to the averaged signals from the DMSO control group. The statistical analysis was performed by one-way ANOVA with post hoc Dunnett's tests, and F, degree of freedom and post hoc P values are indicated in each bar plot. \*\*\*\*P < 0.0001. The post hoc analysis was not performed if the ANOVA tests did not show significance (P > 0.05). The western blot of LC3  $confirmed \, the \, partial \, inhibition \, of \, autophagy \, in \, the \, ATG5-knockdown \, cells.$ f, Similar to e, but in wild-type (WT) or Atg5-knockout (Atg5 KO) mouse embryonic fibroblast lines (MEF) transfected with full-length mHTT (flHTT-Q73). The western blot of LC3 confirmed the inhibition of autophagy in the Atg5-KO cells. For all panels, n indicates the number of independently plated wells, and bars represent mean and s.e.m. Full-blots of cropped gels are shown in Supplementary Fig. 1.



# $\label{prop:pathways} Extended \ Data \ Fig.\ 4\ |\ Potential\ influence\ on\ c-Raf \ and\ KSP\ pathways following\ treatment\ with\ the\ mHTT-LC3\ linker\ compounds.$

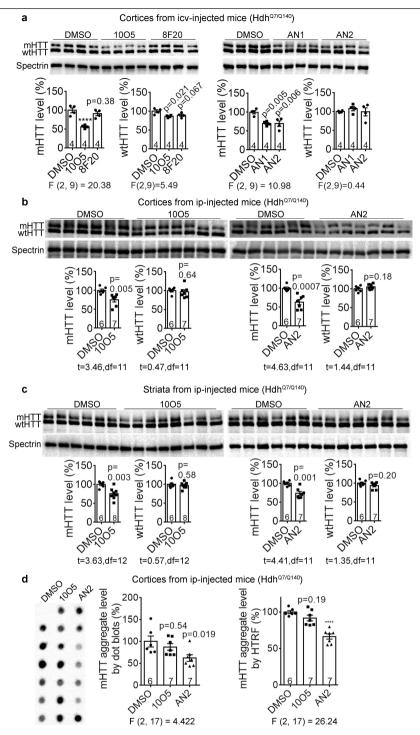
 $\label{eq:approx} \textbf{a}, Representative results (from three biological repeats) of the in vitro c-Raf kinase assay (see Methods) showing that only 1005 inhibits c-Raf activity within the concentration range tested. <math display="block">\textbf{b}, Representative western blots and quantifications of phospho-MEK and phospho-ERK as indicators of Raf activity (left) and phospho-BUBR1 as an indicator of KSP inhibition (right) in cultured cortical neurons treated with indicated compounds (100 nM for 1005, 8F20, AN1, and 50 nM for AN2) or the DMSO control. <math display="block">\textbf{c}, Similar \ to \ \textbf{b}, but \ in$ 

immortalized fibroblasts from a patient with HD (Q47). Note that phospho-BUBR1 is essentially absent and too weak to quantify, indicating that KSP was not inhibited by any of the compounds at the concentration tested. Data are mean  $\pm$  s.e.m. In **b**, **c**, all data were corrected by the loading control ( $\beta$ -tubulin) and normalized to the averaged signal of the DMSO control group. The statistical analysis was performed by one-way ANOVA and F, degree of freedom and post hoc P values are indicated in each bar plot. The n number indicates the number of independently plated and treated wells.



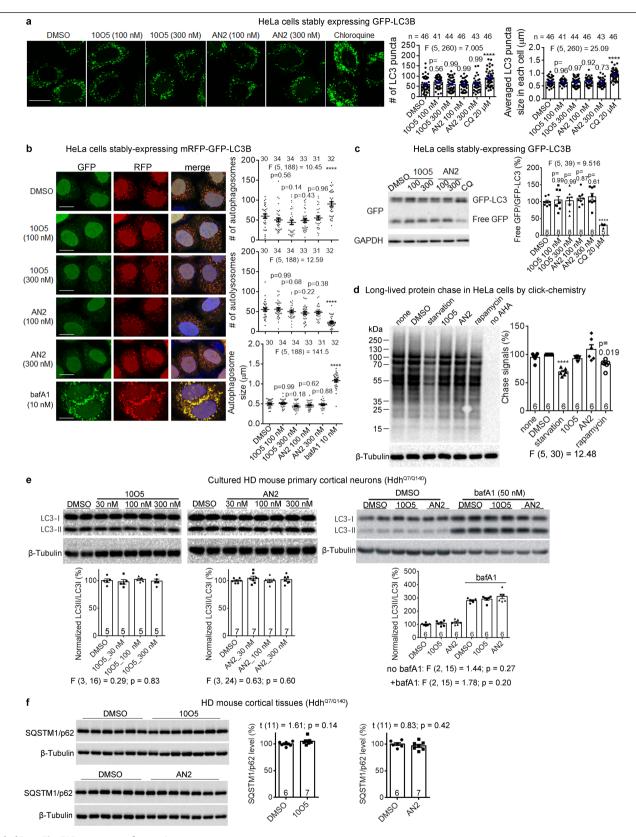
**Extended Data Fig. 5** | **mHTT-LC3 linker compounds lowered mHTT in transgenic HD flies. a**, Overlay between LC3B and predicted Atg8 structure showing high structural similarities. **b**, Transgenic flies expressing full-length HTT(Q128) driven by elav-GAL4 were fed with indicated compounds at 10  $\mu$ M for 6 days, and protein lysates were extracted from the heads. mHTT was then measured by HTRF using the 2B7/MW1 antibody pair. Each dot represents the HTRF signal from each individual sample extracted from five fly heads. All the data were normalized to the average of the DMSO-fed control samples. The

statistical analysis was performed by one-way ANOVA and Dunnett's post hoc tests. F(4,31)=15.67;\*\*\*\*P<0.0001.  $\mathbf{c}$ , 1005 (top) and AN2 (bottom) concentrations in heart plasma and brain tissues were measured by mass spectrometry at the indicated time points for compound-injected mice  $(0.5\,\mathrm{mg\,kg^{-1}})$ . In the brain tissue, the 1005 concentrations were  $-20\,\mathrm{to}$   $-200\,\mathrm{nM}$ , and the AN2 concentrations were  $-20\,\mathrm{to}$   $-40\,\mathrm{nM}$ , close to the effective doses that were capable of lowering mHTT in cultured neurons. Data are mean  $\pm$  s.e.m.



Extended Data Fig. 6 | mHTT-LC3 linker compounds lowered mHTT in vivo in mouse brains. a, Western blots (4 mice (3 months old) for each group) and quantifications of mHTT and wtHTT in the cortices from  $Hdh^{Q7/Q140}$ -knock-in mice with intracerebroventricular injection of the indicated compounds (2  $\mu$ l at 25  $\mu$ M for each mouse) for 10 days at one dose per day. HTT was detected by western blot using the 2166 antibody, and the statistical analysis was performed by one-way ANOVA and post hoc Dunnett's tests. F, degree of freedom and post hoc P values are indicated below each bar plot.  $\mathbf{b}$ , Similar to  $\mathbf{a}$ , except that the compounds were delivered to 5-month-old  $Hdh^{Q7/Q140}$  mice by intraperitoneal injection (0.5 mg kg $^{-1}$ ) for 14 days at one dose per day.  $\mathbf{c}$ , Similar to  $\mathbf{b}$ , but from striata of intraperitoneally injected mice. The mice were injected

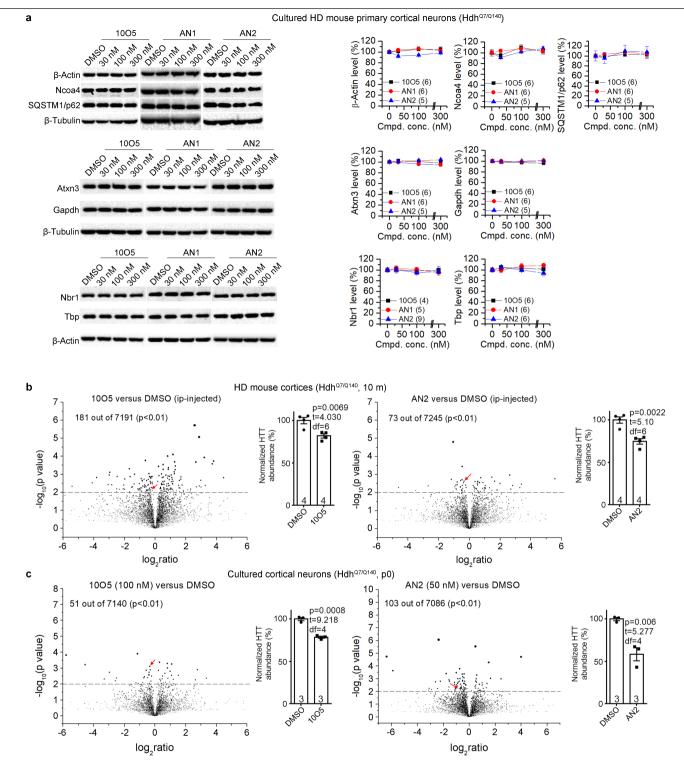
at 10 months old for 14 days at one dose per day.  $\mathbf{d}$ , Left, representative dot blot results (from two technical replicates) of the protein lysates from  $\mathbf{b}$  using the 4C9 antibody, which preferentially detects mHTT aggregates $^{23}$ . Middle, quantification of the dot blots based on the averaged signals from two technical replicates. Right, measurement of mHTT aggregates by the 4C9–4C9 HTRF assay $^{23}$ . In all panels, n indicates the number of mice tested, and bars represent mean and s.e.m. For quantification, two to three technical replicates were averaged for each mouse. Statistical analysis was performed by one-way ANOVA with post hoc Dunnett's tests, and F, degree of freedom and post hoc P values are indicated in each bar plot.



 $\textbf{Extended Data Fig. 7} | See \ next \ page \ for \ caption.$ 

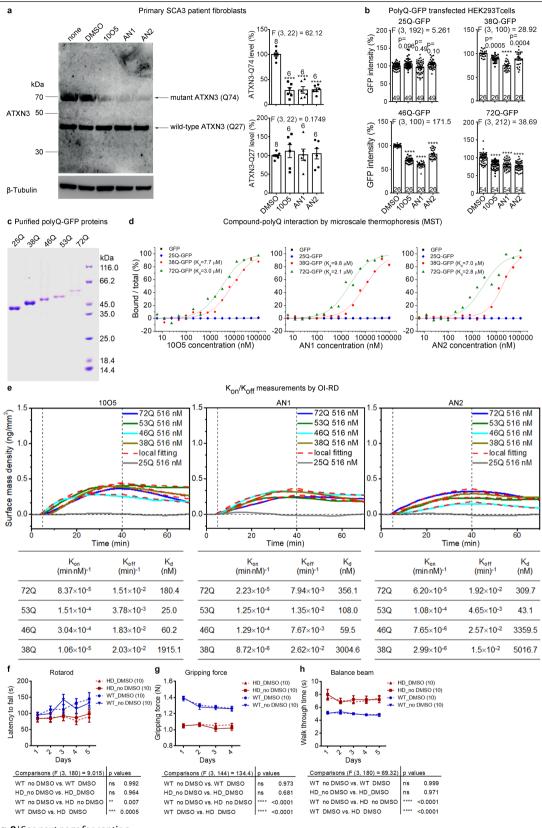
Extended Data Fig. 7 | mHTT-LC3 linker compounds did not influence  $\textbf{autophagy.} \textbf{a}, \text{HeLa cells stably expressing GFP-LC3B} \ were \ treated \ with \ 2\,\mu l$ vehicle (0.1% DMSO), 1005 or AN2 for the indicated concentration for 24 h; chloroquine (CQ, 20 µM) treatment was used as a control. After 24 h, cells were fixed and images were acquired by confocal microscopy. The number and size of GFP vesicles per cell was determined using ImageJ software (n indicated on top of each plot). For each treatment, more than 20,000 puncts were quantified (~100 puncta per cell from 226 cells). Scale bar, 10 μm. b, Representative images and quantifications of the numbers of autophagosomes (GFP+ puncta) and autolysosomes (RFP+GFP- puncta) in HeLa cells stably expressing mRFP-GFP-LC3B. Scale bar, 10 µm. Autophagosome numbers or sizes were not influenced by 1005 and AN2 at the indicated concentrations after 24 h treatment (or 4 h treatment, not shown). The autophagsome fusion was also unaffected as indicated by the autolysosome number. Note that the autophagosome and autolysosome numbers and sizes were based on image analysis of the puncta, some of which may represent  $multiple\,ve sicles.\,Green\,ve sicles\,are\,considered\,to\,be\,autophagosomes\,(GFP^+)$ puncta) and red vesicles are considered to be both autophagosomes and autolysosomes. The number of autolysosomes (RFP+GFP-puncta) was calculated by subtracting the number of green vesicles from that of the red vesicles. More than 10,000 puncta from 194 cells were analysed. c,

Representative western blots and quantifications of HeLa cells stably expressing GFP-LC3B. The 'free GFP' was generated by lysosomal cleavage, and thus the free GFP/GFP-LC3B ratio was used as an index for autophagy flux, which was unaffected by 1005 or AN2, but decreased by the autophagy flux inhibitor chloroquine. d, Representative western blots and quantifications of the chase signal of long-lived proteins in HeLa cells as an indicator of autophagy flux (see Methods). Consistent with previous reports<sup>33</sup>, starvation reduced the long-lived protein chase signal, whereas rapamycin treatment had a milder effect. The mHTT-LC3 linker compounds 1005 and AN2 had no influence in this assay. e, Representative western blots and quantifications of LC3 in cultured cortical neurons treated with the indicated compounds. Normalized LC3-II/ LC3-I was used as the indicator of autophagy. Right blot: 1005, 100 nM; AN2, 50 nM. f, SQSTM1 (p62) levels were determined by western blot for the cortical tissues from mice injected with the indicated compounds or DMSO control. Bars indicate mean and s.e.m.; n indicated in each bar shows the number of cells  $(\mathbf{a}, \mathbf{b})$ , the number of independently plated wells  $(\mathbf{c} - \mathbf{e})$  or the number of mice  $(\mathbf{f})$ . Data are mean ± s.e.m. The statistical analysis was performed by one-way ANOVA with post hoc Dunnett's tests ( $\mathbf{a}-\mathbf{e}$ ) or two-tailed unpaired t-tests ( $\mathbf{f}$ ). Note that the post hoc tests were not performed if the ANOVA tests failed to show significance. \*\*\*\*P < 0.0001 (post hoc test).



**Extended Data Fig. 8** | **Investigation on the specificity of mHTT-lowering effects of mHTT-LC3 linker compounds. a**, Representative western blots and quantifications of cultured cortical neurons treated with the indicated compounds. None of the proteins tested showed a clear effect (>10%). **b**, Volcano plots of the proteomics analysis of cortices from intraperitoneally injected HD mice (10 month old, 4 mice per group, injected for 14 days). Mice were injected with  $0.5\,\mathrm{mg\,kg^{-1}}$  protein with  $110\,\mathrm{\mu g\,kg^{-1}}$  DMSO, and equal amount of vehicle containing DMSO was injected in the control mice. Only proteins

detected in both groups of samples used for comparisons were calculated and plotted. Red arrows indicate HTT. See Supplementary Table 2 for complete datasets. The bar plots indicate the total HTT levels normalized to the DMSO control. The actual mHTT reduction is anticipated to be higher, because the compounds reduced mHTT in an allele-selective manner.  ${\bf c}$ , Similar to  ${\bf b}$ , but in cultured cortical neurons (from postnatal day 0 pups, three wells per group). See Supplementary Table 3 for complete datasets. In all panels, data are mean  $\pm$  s.e.m.



 $\textbf{Extended Data Fig. 9} \, | \, \textbf{See next page for caption}.$ 

# $\label{lem:extended} Extended \ Data \ Fig. \ 9 \ | \ mHTT-LC3 \ linker \ compounds \ lowered \ the \ mutant \ ATXN3 \ protein \ with \ polyQ \ expansion \ in \ an \ allele-selective \ manner.$

**a**, Representative western blots and quantifications of ATXN3 levels in a fibroblast line from a patient with SCA3 treated with the indicated compounds. The lowering of mutant (Q74) but not wild-type (Q27) ATXN3 was observed by treatment of linker compounds tested. **b**, Quantification of the GFP intensity as an indicator of polyQ-sfGFP (25Q-GFP, 38Q-GFP, 46Q-GFP and 72Q-GFP) protein levels in transfected HEK293T cells treated with the indicated compounds using Incucyte. Reduction of 72Q-GFP, 46Q-GFP and 38Q-GFP but not 25Q-GFP was observed. In **a** and **b**, the compound concentrations were 100 nM for 1005 and AN1, and 50 nM for AN2. Bar plots present mean  $\pm$  s.e.m., and n indicates the number of independently plated wells. **c**, SDS-PAGE analysis of polyQ-sfGFP proteins (25Q, 38Q, 46Q, 53Q and 72Q) purified from HEK293T cells. The protein purification methods were similar to those for HTT proteins. **d**, Binding of 1005, AN1 and AN2 to sfGFP (GFP) or different polyQ-

sfGFP (25Q–GFP, 38Q–GFP and 72Q-GFP) proteins in standard treated capillaries measured by MST, performed and analysed similarly as in Extended Data Fig. 2b. All these compounds interact with 38Q–GFP and 72Q–GFP but not with 25Q–GFP or GFP. **e**, Association–dissociation curves of surface-immobilized compounds 10O5, AN1 and AN2 with polyQ–sfGFP (72Q, 53Q, 46Q, 38Q and 25Q) proteins. For all association–dissociation curves, vertical dashed lines mark the starts of association and dissociation phases of the binding event. The red dashed curves are fits to a Langmuir reaction model with the fitting parameters listed at the bottom of each plot. No binding signals were observed for 25Q–sfGFP (25Q). **f–h**, Results of mouse behavioural test performed similarly to those in Fig. 5d–f, except that the mice were injected with saline (0.9% NaCl) with DMSO (110  $\mu$ g kg $^{-1}$ ) or without DMSO. The statistical analysis was performed by two-way ANOVA with post hoc Bonferroni's tests, and *F*, *P* values and degrees of freedom are indicated in the table below each plot. In all panels, data are mean  $\pm$  s.e.m.

#### Extended Data Table 1 | Summary of data on mHTT lowering or rescue of HD-relevant phenotypes

Model and treatment	Readout and figures	Compound effects
		10O5: 26.0±3.3% lowering
cultured primary cortical	the mHTT level by Western-blot (Fig.	8F20: 40.1±12.6% lowering
neurons, from mice (Hdh <sup>Q7/Q140</sup> )	2a&d)	AN1: 35.7±2.8% lowering
		AN2: 34.0±6.2% lowering
		10O5: 45.1±4.0% lowering
primary human HD patient	the mHTT level by HTRF (Fig. 3a)	8F20: 44.8±4.9 lowering
fibroblasts (Q49)	and min in level by it it is (ing. ea)	AN1: 46.1±6.2% lowering
		AN2: 54.8±7.5% lowering
		10O5: 34.3±5.4% lowering
primary human HD patient	the mHTT level by HTRF (Fig. 3a)	8F20: 28.7±3.6% lowering
fibroblasts (Q55)	( (	AN1: 26.5±7.0% lowering
		AN2: 39.3±4.8% lowering
		10O5: 20.9±2.7% lowering
primary human HD patient	the mHTT level by HTRF (Fig. 3a)	8F20: 22.9±5.3% lowering
fibroblasts (Q68)	, (3, /	AN1: 26.4±2.8% lowering
		AN2: 18.1±5.1% lowering
		10O5: 31.4±3.2% lowering
	the mHTT level by HTRF (Ext. Data.	8F20: 28.9±3.7% lowering
	Fig. 5c)	AN1: 39.3±3.2% lowering
HD patient iPSC-derived		AN2: 40.5±2.8% lowering
neurons (Q47)	surface area of each neuron by Tuj1	10O5: 69.5±1.6% rescue
		8F20: 51.6±3.1% rescue
	staining (Fig. 5a)	AN1: 58.8±4.9% rescue
		AN2: 64.4±2.7% rescue
		10O5: 30.2±4.5% lowering
immortalized human HD patient	the mHTT level by HTRF (Fig. 3b)	8F20: 22±4.8% lowering
fibroblasts (Q47)	, , ,	AN1: 42.0±3.9% lowering
		AN2: 41.4±5.1% lowering
		10O5: 43.3±2.2% lowering
icv-injected mice (Hdh <sup>Q7/Q140</sup> )	the mHTT level by Western-blot (Ext.	8F20: 9.1±5.3% lowering (n.s.)
,	Data Fig. 9a)	AN1: 29.9±2.9% lowering
		AN2: 30.3±7.4% lowering
	the cortical mHTT by Western-blot	10O5: 24.8±4.2% lowering
	(Ext. Data Fig. 9b)	AN2: 36.6±7.4% lowering
	the striatal mHTT by Western-blot	10O5: 22.9±2.3% lowering
	(Ext. Data Fig. 9c)	AN2: 26.3±5.5% lowering
	the cortical HTT by MASS-SPEC	10O5: 18.1±2.4% lowering
ip-injected mice (Hdh <sup>Q7/Q140</sup> )	(Ext. Data Fig. 11b)	AN2: 25.2±3.2% lowering
p-injected files (Fidin )	latency to fall by rotarod tests (Fig.	10O5: (60.8% averaged rescue)
	5d)	AN2: (64.3% averaged rescue)
	passing time by balance beam tests	10O5: (77.2% averaged rescue)
	(Fig. 5e)	AN2: (92.8% averaged rescue)
	grinning force tosts (Fig. Ef)	10O5: (43.6% averaged rescue)
	gripping force tests (Fig. 5f)	AN2: (52.4% averaged rescue)

A summary table showing the percentage lowering of mHTT or HTT levels, and the percentage rescue of HD-relevant phenotypes (normalized to the difference between HD and wild type) in different HD models assayed by different approaches under optimal conditions. The corresponding data are indicated in the middle column. The percentage change/rescue is presented as mean + s.e.m.



Corresponding	author(s	s): Boxun L	.u, Yiyan	Fei, Yu [	Din
---------------	----------	-------------	-----------	-----------	-----

# **Reporting Summary**

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main

## Statistical parameters

text	, or N	Methods section).
n/a	Cor	nfirmed
	$\boxtimes$	The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
	$\boxtimes$	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
		The statistical test(s) used AND whether they are one- or two-sided Only common tests should be described solely by name; describe more complex techniques in the Methods section.
	$\boxtimes$	A description of all covariates tested
	$\boxtimes$	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
		A full description of the statistics including <u>central tendency</u> (e.g. means) or other basic estimates (e.g. regression coefficient) AND <u>variation</u> (e.g. standard deviation) or associated <u>estimates of uncertainty</u> (e.g. confidence intervals)
		For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i> ) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted <i>Give P values as exact values whenever suitable.</i>
X		For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
$\boxtimes$		For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
	$\boxtimes$	Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated
		Clearly defined error bars State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on statistics for biologists may be useful.

#### Software and code

Policy information about availability of computer code

All data collection softwares came with the equipment utilized for experiments, including IncuCyte 2011A, MO. Control (NT.115), Data collection

PerkinElmer EnVision Manager Version 1.13, HKL2000 V719, Image Lab Version 3.0, ZEN 2.3

The softwares utilized for analysis were all commercially available or could be downloaded from open source, including GraphPad Prism Data analysis 7, ImageJ 1.52a, Origin8, Microsoft Excel 2016, PASS 16, Nanotemper analysis (1.5.41), PyMOL 2.2, HKL2000, Phaser 2.8, Mascot 2.3, PASS 16, IncuCyte 2011A, MO.Affinity Analysis (NT.115).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

#### Data

Policy information about availability of data

All manuscripts must include a <u>data availability statement</u>. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The protein structure data has been uploaded to the PDB database with entry number 6J04. The source data in excel files have been provided for all essential plots in the figures. The full gel blots and the proteomics datasets have been provided as supplementary tables. All the other data are available from the authors upon request.

Field-specific reporting						
•	est fit for your research. If you are not sure, read the appropriate sections before making your selection.					
X Life sciences	Behavioural & social sciences Ecological, evolutionary & environmental sciences					
	the document with all sections, see <a href="mailto:nature.com/authors/policies/ReportingSummary-flat.pdf">nature.com/authors/policies/ReportingSummary-flat.pdf</a>					
Life scier	nces study design					
All studies must dis	sclose on these points even when the disclosure is negative.					
Sample size	To ensure to reach a statistical power>0.8, power analyses were performed for each assay based on estimated values by PASS 16 (https://www.ncss.com/software/pass/) before experiments. Estimation was based on our previously published results on similar experiments and preliminary experiments. The power analysis suggested n > =3 for mHTT level measurements and n > =5 for behavioral experiments. In all the experiments we performed, we have used a larger n than these numbers, and we also performed post-experiment power analyses to ensure that power > 0.8 for all the significant differences.					
Data exclusions	Data were not excluded unless clear experimental failures occurred, including cell contamination, gel transfer failures and lack of signals in positive controls. The exclusion criteria were pre-established.					
Replication	All experimental data was reliably reproduced in multiple independent experiments as indicated in the figure legends. The protein-compound interaction experiments and the HTT-lowering experiments have been replicated by at least two independent researchers.					
Randomization	For the in vivo experiments in the mouse, randomization was performed by assigning random numbers. For the Drosophila experiments, the flies were randomly distributed into the vials after anesthesia.					
Blinding	As indicated in the figure legends, the investigators were blinded during data collection and analysis where possible, included immunofluorescence experiments, drug treatment experiments in mouse and fly models for HTT level measurements and behavioral assays.					
·	g for specific materials, systems and methods					
	aterials & experimental systems Methods					
n/a Involved in th	ne study n/a   Involved in the study    ChIP-seq					
Antibodies						
Eukaryotic	cell lines MRI-based neuroimaging					
	Palaeontology					
Animals and other organisms						
∐  X  Human res	Human research participants					

# Unique biological materials

Policy information about <u>availability of materials</u>

Obtaining unique materials All unique materials (some of the patient fibroblast lines) are readily available from the authors.

# **Antibodies**

Antibodies used

Anti-β-tubulin (Abcam, cat. no. ab6046, lot no. GR3209100-1, 1:10000)

Anti-TUBB3 (Biolegends (Covance), cat. no. 801202, clone TUJ1, lot no. B2335555, 1:500)

Anti-ATXN3 (Millipore, cat. no. MAB5360, clone no. 1H9, lot no. 3096481,1:1000):

Anti-Gapdh (Proteintech. cat. no. 60004-1, lot no. 10004129, 1:5000)

Anti-NBR1 (ThermoFisher Scientific, cat. no. PA5-54660, lot no. SJ2462971A, 0.4 µg/mL)

Anti-β-actin (Beyotime, cat.no. AA128, clone no.AC-74, lot no. 031918180821, 1:1000)

Anti-TBP antibody (Abcam, cat. no. ab818, clone no.1TBP18, lot no. GR315577-3, 1:1000)

Anti-SQSTM1 antibody (ThermoFisher Scientific, cat.no. PA5-27247, lot no. SC2360851N, 1:1000)

Anti-spectrin antibody (Millipore, cat.no. MAB1622, clone no. AA6, lot no. 2943221, 1:2000)

Anti-Ncoa4 antibody (Santa cruz, cat.no. sc-373739, clone no. C-4, 1:200)

Anti-GST antibody (ProteinTech, cat.no. HRP-66001, clone no. 3G12B10, lot no. 20000091, 1:2000)

anti-MBP antibody (ProteinTech, cat.no. 15089-1-AP, lot no. 00058716, 1:3000)

Anti-His antibody (Beyotime, cat. no. AH367, clone no.AD1.1.10, lot no.011018180312, 1:100)

Anti-LC3B antibody (ThermoFisher Scientific, cat.no.PA1-16930, lot no.T12629311, 1:1000; ThermoFisher Scientific, cat.no. 700712, clone no. 2H30L32, lot no. 2086347, 1:200)

Phospho-Erk1/2 Pathway Sampler Kit (Cell Signaling Technology cat.no.9911, 1:1000; Phospho-MEK1/2, clone no.41G9, lot no.18, Phospho-p44/42 MAPK, clone no.D13.14.4E, lot no.24)

Anti-GFP (Cell Signaling Technologies, cat. no. 2956, clone no. D5.1, lot no. 4, 1:1000)

Anti-BubR1 antibody (BD Transduction, cat. no. 612503, clone no. 9/BUBR1, 1:1000)

Anti-Huntingtin Protein antibody 2166 (Millipore, cat. no. MAB2166, clone no.1HU-4C8, lot no. 2943221, 1:1000)

Anti-polyQ antibody 3B5H10 (Sigma, cat. no. P1874, clone no. 3B5H10, lot no. 047M4820V, 1:1000)

Anti-HTT antibody (D7F7)XP (Cell Signaling Technologies, cat. no. 5656s, clone no. D7F7, lot no. 4, 1:1000)

The other HTT antibodies including 2B7, ab1 and MW1 were previously published and characterized by other groups, and they originally obtained from those groups. They were diluted at 1:1000 for Western-blots.

Validation

The anti-β-tubulin antibody has been validated for Western-blots of both human and mouse samples by many previous publications (e.g. PMID: 28869595). The anti-TUBB3 antibody has been validated for immunocytochemistry of human sample by many previous publications (e.g. PMID: 30545851). The anti-ATXN3 antibody has been validated for Western-blots of both human and mouse samples by many several publications (e.g. PMID: 28180282). The anti-Gapdh antibody has been validated for Western-blots of mouse sample by many previous publications (e.g. PMID: 31091447). The anti-NBR1 antibody has been validated for Western-blots of mouse in antibodypedia (https://www.antibodypedia.com/gene/8116/NBR1/antibody/3592759/ PA5-54660). Anti-β-actin antibody has been validated for Western-blots of both human and mouse samples by many previous publications (e.g. PMID: 30459625). The anti-TBP antibody has been validated for Western-blots of both human and mouse samples by many previous publications (e.g. PMID: 28280206). The anti-SQSTM1 antibody has been validated for Western-blots of mouse sample by many previous publications (e.g. PMID: 28869595). The anti-spectrin antibody has been validated for Western-blots of both human and mouse samples by many previous publications (e.g. PMID: 28869595). The anti-Ncoa4 antibody has been validated for Western-blots of both human and mouse samples by many previous publications (e.g. PMID: 30630985). The anti-GST antibody and the anti-MBP antibody has been validated for in vitro pull-down and Western-blot by experimental data in this study (Fig 4a-b). Anti-His antibody has been validated for immunostaining by experimental data in this study (Fig 4c). The anti-LC3B antibody (ThermoFisher Scientific, cat.no. PA1-16930) has been validated for Western-blots of mouse sample by vender (https://www.thermofisher.com/cn/zh/antibody/product/LC3B-Antibody-Polyclonal/PA1-16930). The anti-LC3B antibody (ThermoFisher Scientific, ThermoFisher Scientific, cat.no. 700712) has been validated by previous publications for immunostaining in human and mouse cells (PMID: 29151587, 22622129). The anti-GFP antibody has been validated by previous literature (PMID: 31112137; PMID: 31067454; PMID: 30996031). The Phospho-Erk1/2 Pathway Sampler Kit have been validated for Western-blots of mouse sample by several previous publications (e.g. PMID: 29636449). The anti-BubR1 antibody has been validated for Western-blots of both human and mouse samples by many previous publications (e.g. PMID: 27528194). The anti-HTT antibody (D7F7)XP have been validated for Western-blots of both human and mouse samples by many previous publications.(e.g. PMID: 26863614, PMID: 23575829). The other HTT antibodies including 2166, 3B5H10, 2B7, ab1 and MW1 have been validated for Western-blots of both human and mouse samples by many previous publications (e.g. PMID: 25738228). The antibody 2166 has also been validated for immunostaining experiments in mouse samples by this study (Fig. 4d).

# Eukaryotic cell lines

Policy information about  $\underline{\mathsf{cell\ lines}}$ 

Cell line source(s)

Some of the primary patient fibroblasts were obtained from HD patients (Q47, Q49, Q55) and healthy sibling (WT, Q19) controls in a Mongolian Huntington's disease family. The HD Q68 fibroblast line was obtained from Coriell Cell Repositories (Camden, NJ, USA). The PD line was obtained from an idiopathic Parkinson's disease patient, and the SCA3 line was obtained from a SCA3 patient with the ATXN3 expansion mutation (Q74). The studies were approved by The Ethic Community of Institutes of Biomedical Sciences at Fudan University (#28) for obtaining the HD and wild-type patient fibroblasts, and by Huashan Hospital Institutional Review Board at Fudan University (#174) for obtaining the PD and SCA3 patient fibroblasts. Verbal and written consent was obtained from patients. The procedures were in compliance with all relevant ethical regulations. The immortalized fibroblasts were generated by infection of lentivirus expressing SV40T. For generation of iPS cells (iPSCs), the primary fibroblasts were transduced with the retroviral STEMCCA polycistronic reprogramming system (Millipore, cat. no. SCR548). The iPSCs were confirmed positive for Tra-1-81, Tra-1-60, SSEA-4 and Nanog by immunofluorescence and flow-cytometry. All four vector-encoded transgenes were found to be silenced and the karyotype was normal. iPSC were cultured in E8 medium (ThermoFisher Scientific, cat. no. A1517001) on Matrigel (Corning, cat. no. 354277) surface. iPSCs were differentiated to Pax6-expressing primitive neuroepithelia (NE) for 10-12 days in a neural induction medium. Sonic hedgehog (SHH, 200 ng/ml) was added at days 10-25 to induce ventral progenitors. For neuronal differentiation, neural progenitor clusters were dissociated and placed onto poly-ornithine/laminin-coated coverslips at day

26 in Neurobasal medium (ThermoFisher Scientific, cat. no. 21103049), with 1× B-27 (ThermoFisher Scientific, cat. no. 17504044), 1× N-2 (ThermoFisher Scientific, cat. no. 17504048), brain derived neurotrophic factor (BDNF, 20 ng/ml, Protech, cat. no. 450-02), glial-derived neurotrophic factor (GDNF, 10 ng/ml, Protech, cat. no. 450-10), insulin-like growth factor 1 (IGF1, 10 ng/ml, Protech, cat. no. 100-11) and Vitamin C (Sigma cat. no. D-0260, 200 ng/ml). The mouse striatal cells (STHdh) were obtained from Coriell Cell Repositories (Camden, NJ, USA). The HEK293T cells and the HeLa cells were originally obtained from American Type Culture Collection (ATCC). Atg5 WT and KO MEFs were from N. Mizushima.

Authentication

The HEK293T and HeLa cell lines were authenticated by Short Tandem Repeat (STR) profiling methods. The Atg5 WT and KO MEFs were obtained directly from the laboratory which generated these cell lines (N. Mizushima), and they were further authenticated by Short Tandem Repeat (STR) profiling methods comparing with primary cultured MEFs. The patient fibroblasts were obtained and cultured from patients, and they were not authenticated.

Mycoplasma contamination

The cells were tested every two months by a TransDetect PCR Mycroplasma Detection Kit (Transgen Biotech, cat. no. FM311-01) to ensure that they are mycoplasma free.

Commonly misidentified lines (See ICLAC register)

HeLa cells were used in the HTT-LC3 colocalization experiments, because it is commonly used cell line for autophagy experiments and it showed more distinct LC3 puncta than other cells that we have tested. In addition, it has high transfection efficiency.

# Animals and other organisms

Policy information about <u>studies involving animals</u>; <u>ARRIVE guidelines</u> recommended for reporting animal research

Laboratory animals

The fruitfly experiments used Drosophila Melanogaster, and adult virgin female flies were used for experiments at indicated days of age (ranging from 0 to 50 days old after eclosion). The mouse experiments used the C57BL/6 strain including both male and female of the desired genotype. For icv experiments, the age was 3 months +/- 0 days old; for ip-injection followed by testing of cortical HTT, the age was 5 months +/- 0 days old; for ip-injection followed by testing of striatal HTT, the age was 10 months +/- 3 days old; for ip-injection followed by behavioral analysis, the age was 10 months +/- 3 days old.

Wild animals

The study did not involve wild animals.

Field-collected samples

The study did not involve samples collected from the field.

# Human research participants

Policy information about studies involving human research participants

Population characteristics

Since we are testing the compounds in patient cells rather than patient groups, we do not have population characteristics of human participants. For each patient, we cultured many cells from them and treat different group of cells with different compounds to test their effects within each patient cell lines.

In general, the patients were diagnosed base on symptoms and genetic testings, and they received no treatment at the time they provided dermal fibroblasts. The HD patients (Q47/Q19, 46-year-old female; Q49/Q19, 44-year-old male; Q55/Q19, 39-year-old male) and the healthy sibling control (Q19/Q19, 39-year-old female) were from a Mongolian family. The The SCA3 patient was a 32 years old female patient when providing the dermal fibroblasts. She first came to the clinic complaining with clumsy and slowness in the lower limbs and left upper limb for 1 year. Her father and grandfather had the same symptoms but had passed away. After a one year follow up, she developed unstable walking. She was diagnosed as spinocerebellar ataxia and confirmed by genetic testing with the repeat number of Q74/Q27 in ATXN3.

The PD patient was a 74 years old male patient when providing the dermal fibroblasts. He developed tremor, rigidity and bradykinesia for 6 years. The symptoms started at age 68 and he was diagnosed as PD at age 69 and followed up in our center for 5 years. A panel containing 254 PD and related genes and PD MLPA were carried out in the patient but did not find any known mutations related to PD. DAT-PET CT found the decreased DAT binding in the right caudate nucleus and putamen.

Recruitment

The HD and SCA3 patients were recruited by clinical symptoms and confirmed with genetic testing. The PD patients were recruited by clinical symptoms and confirmed by follow-up visits of more than 5 years and DAT PECT-CT. The recruitment could be biased because only a few patients who see the collaborating doctor and want to donate dermal fibroblasts for potential future research were selected. This is typical for preclinical studies, and our study is comparing different compound treated groups within each of the cell line, and thus not influenced by patient-to-patient variations. Nonetheless, while we have tested multiple patient cells and obtained consistent results, it is still possible that some of the other patient cells show different results.

# Pan-cancer whole-genome analyses of metastatic solid tumours

https://doi.org/10.1038/s41586-019-1689-y

Received: 9 September 2018

Accepted: 20 September 2019

Published online: 23 October 2019

Open access

Peter Priestley<sup>1,2,12</sup>, Jonathan Baber<sup>1,2,12</sup>, Martijn P. Lolkema<sup>3,4</sup>, Neeltje Steeghs<sup>3,5</sup>, Ewart de Bruijn<sup>1</sup>, Charles Shale<sup>2</sup>, Korneel Duyvesteyn<sup>1</sup>, Susan Haidari<sup>1,3</sup>, Arne van Hoeck<sup>6</sup>, Wendy Onstenk<sup>1,3,4</sup>, Paul Roepman<sup>1</sup>, Mircea Voda<sup>1</sup>, Haiko J. Bloemendal<sup>7,8</sup>, Vivianne C. G. Tjan-Heijnen<sup>9</sup>, Carla M. L. van Herpen<sup>8</sup>, Mariette Labots<sup>10</sup>, Petronella O. Witteveen<sup>11</sup>, Egbert F. Smit<sup>3,5</sup>, Stefan Sleijfer<sup>3,4</sup>, Emile E. Voest<sup>3,5</sup> & Edwin Cuppen<sup>1,3,6\*</sup>

Metastatic cancer is a major cause of death and is associated with poor treatment efficacy. A better understanding of the characteristics of late-stage cancer is required to help adapt personalized treatments, reduce overtreatment and improve outcomes. Here we describe the largest, to our knowledge, pan-cancer study of metastatic solid tumour genomes, including whole-genome sequencing data for 2,520 pairs of tumour and normal tissue, analysed at median depths of 106× and 38×, respectively, and surveying more than 70 million somatic variants. The characteristic mutations of metastatic lesions varied widely, with mutations that reflect those of the primary tumour types, and with high rates of whole-genome duplication events (56%). Individual metastatic lesions were relatively homogeneous, with the vast majority (96%) of driver mutations being clonal and up to 80% of tumour-suppressor genes being inactivated bi-allelically by different mutational mechanisms. Although metastatic tumour genomes showed similar mutational landscape and driver genes to primary tumours, we find characteristics that could contribute to responsiveness to therapy or resistance in individual patients. We implement an approach for the review of clinically relevant associations and their potential for actionability. For 62% of patients, we identify genetic variants that may be used to stratify patients towards therapies that either have been approved or are in clinical trials. This demonstrates the importance of comprehensive genomic tumour profiling for precision medicine in cancer.

In recent years, several large-scale whole-genome sequencing (WGS) analysis efforts have yielded valuable insights into the diversity of the molecular processes that drive different types of adult<sup>1,2</sup> and paediatric<sup>3,4</sup> cancer and have fuelled the promises of genome-driven oncology care<sup>5</sup>. However, most analyses were done on primary tumour material, whereas metastatic cancers—which cause the bulk of the disease burden and 90% of all cancer deaths—have been less comprehensively studied at the whole-genome level, with previous efforts focusing on tumour-specific cohorts <sup>6-8</sup> or at a targeted gene panel <sup>9</sup> or exome level<sup>10</sup>. As cancer genomes evolve over time, both in the highly heterogeneous primary tumour mass and as disseminated metastatic cells<sup>11,12</sup>, a better understanding of metastatic cancer genomes will be highly valuable to improve on adapting treatments for late-stage cancers.

Here we describe the pan-cancer whole-genome landscape of metastatic cancers based on 2,520 paired tumour ( $106 \times \text{average depth}$ ) and normal (blood,  $38 \times$ ) genomes from 2,399 patients (Supplementary Tables 1 and 2, Extended Data Fig. 1). The sample distribution over age

and primary tumour types broadly reflects the incidence of solid cancers in the Western world, including rare cancers (Fig. 1a). Sequencing data were analysed using an optimized bioinformatic pipeline based on open source tools (Methods, Supplementary Information) and identified a total of 59,472,629 single nucleotide variants (SNVs), 839,126 multiple nucleotide variants (MNVs), 9,598,205 insertions and deletions (indels) and 653,452 structural variants (SVs) (Supplementary Table 2).

#### **Mutational landscape of metastatic cancer**

We analysed the mutational burden of each class of variant per cancer type based on the tissue of origin (Fig. 1, Supplementary Table 2). In line with previous studies on primary cancers<sup>13,14</sup>, we found extensive variation in the mutational load of up to three orders of magnitude both within and across cancer types.

The median SNV counts per sample were highest in skin, predominantly consisting of melanoma (44,000) and lung (36,000) tumours,

<sup>1</sup>Hartwig Medical Foundation, Amsterdam, The Netherlands. <sup>2</sup>Hartwig Medical Foundation Australia, Sydney, New South Wales, Australia. <sup>3</sup>Center for Personalized Cancer Treatment, Rotterdam, The Netherlands. <sup>4</sup>Erasmus MC Cancer Institute, Rotterdam, The Netherlands. <sup>5</sup>Netherlands Cancer Institute/Antoni van Leeuwenhoekhuis, Amsterdam, The Netherlands. <sup>6</sup>Center for Molecular Medicine and Oncode Institute, University Medical Center Utrecht, Utrecht, The Netherlands. <sup>7</sup>Meander Medisch Centrum, Amersfoort, The Netherlands. <sup>8</sup>Radboud University Medical Center, Nijmegen, The Netherlands. <sup>9</sup>Maastricht University Medical Center, Maastricht, The Netherlands. <sup>10</sup>VU Medical Center, Amsterdam, The Netherlands. <sup>11</sup>Cancer Center, University Medical Center Utrecht, Utrecht, The Netherlands. <sup>12</sup>These authors contributed equally: Peter Priestley, Jonathan Baber. \*e-mail: e.cuppen@hartwigmedicalfoundation.nl

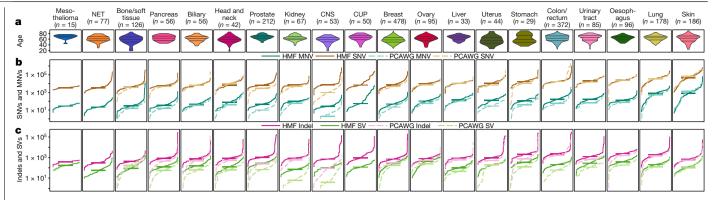


Fig. 1 | Mutational load of metastatic cancer. a, Violin plot showing age distribution of each tumour type, with twenty-fifth, fiftieth and seventy-fifth percentiles marked. b, c, Cumulative distribution function plot (individual samples were ranked independently for each variant type) of mutational load for each tumour type for SNVs and MNVs (b) and indels and SVs (c). The median for

each tumour type is indicated by a horizontal bar. Dotted lines indicate the mutational loads in primary cancers from the PCAWG cohort<sup>14</sup>. Only tumour types with more than ten samples are shown (n = 2,350 independent patients), and are ranked from the lowest to the highest overall SNV mutation burden (TMB), CUP, cancer of unknown primary.

with tenfold higher SNV counts than sarcomas (4,100), neuroendocrine tumours (NETs) (3,500) and mesotheliomas (3,400). SNVs were mapped to COSMIC mutational signatures and were found to broadly match the patterns described in previous cancer cohorts per cancer type<sup>13</sup> (Extended Data Figs. 2, 3). However, several broad spectrum signatures such as S3, S8, S9 and S16 as well as some more specific signature (for example, S17 in specific tumour types) appear to be overrepresented in our cohort. These observations may indicate enrichment of tumours that are deficient in specific DNA repair processes (S3), increased hypermutation processes (S9) among advanced cancers, or reflect the mutagenic effects of previous treatments<sup>15</sup>.

The variation for MNVs was even greater, with lung (median of 821) and skin (median of 764) tumours having five times the median MNV counts of any other tumour type. This can be explained by the wellknown mutational effect of UV radiation (CC>TT) and smoking (CC>AA) mutational signatures, respectively (Extended Data Fig. 2). Although only dinucleotide substitutions are typically reported as MNVs, 10.7% of the MNVs involve three nucleotides and 0.6% had four or more nucleotides affected.

Indel counts were typically tenfold lower than SNVs, with a lower relative rate for skin and lung cancers (Fig. 1c). Genome-wide analysis of indels at microsatellite loci identified 60 samples with microsatellite instability (MSI) (Supplementary Table 2), which represents 2.5% of all tumours (Extended Data Fig. 4). Notably, 67% of all indels in the entire cohort were found in the 60 MSI samples, and 85% of all indels in the cohort were found in microsatellites or short tandem repeats. The highest rates of MSI were observed in central nervous system (CNS) (9.4%), uterine (9.1%) and prostate (6.1%) tumours. For metastatic colorectal cancer lesions, we found an MSI frequency of only 4.0%, which is lower than that reported for primary colorectal cancer, and in line with better prognosis for patients with localized MSI colorectal cancer, which metastasizes less often16.

The median rate of SVs across the cohort was 193 per tumour, with the highest median counts observed in ovarian (412) and oesophageal (372) tumours, and the lowest in kidney tumours (71) and NETs (56). Simple deletions were the most commonly observed subtype of SV (33% of all SVs), and were the most prevalent in every cancer type except stomach and oesophageal tumours, which were highly enriched in translocations (Extended Data Fig. 2).

To gain insight into the overall genomic differences between primary and metastatic cancer, we compared the mutational burden in the Hartwig Medical Foundation (HMF) metastatic cohort with the Pancancer Analysis of Whole Genomes (PCAWG) dataset<sup>14</sup>, which, to our knowledge, is the largest comparable whole-genome sequenced tumour cohort (n=2,583) available so far, and which has 95% of biopsies taken from treatment-naive primary tumours. In general, the SNV mutational load does not seem to be indicative for disease progression as it is not significantly different in this study compared with the PCAWG for most cancer types (Fig. 1b). Prostate and breast cancer are clear exceptions with structurally higher mutational loads ( $q < 1 \times 10^{-10}$ , Mann-Whitney test), which potentially reflects relevant tumour biology and is, for prostate cancer, consistent with other reports<sup>8,17</sup>. CNS tumours also have a higher mutational load that is explained by the different age distributions of the cohorts.

By contrast, the mutational loads of indels, MNVs and SVs are significantly higher across nearly all cancer types analysed (Fig. 1c). This is most notable for prostate cancer, in which we observe a more than fourfold increased rate of MNVs, indels and SVs. Although these observations may represent the advancement of disease and the higher rate of certain mutational processes in metastatic cancers, they are also partially due to differences in sequencing depth and bioinformatic analysis pipelines (Extended Data Figs. 5, 6, Supplementary Information).

#### Copy number alteration landscape

Pan-cancer, the most highly amplified regions in our metastatic cancer cohort contain established oncogenes such as EGFR, CCNE1, CCND1 and MDM2 (Fig. 2). The chromosomal arms 1q, 5p, 8q and 20q are also highly enriched in moderate amplification across the cohort, with each affecting more than 20% of all samples. For amplifications of 5p and 8q, this is probably related to the common amplification targets of TERT and MYC, respectively. However, the targets of amplifications on 1q, which are predominantly found in breast cancers (more than 50% of samples), and amplifications on 20q, which are predominantly found in colorectal cancers (more than 65% of samples), are less clear.

Overall, an average of 23% of the autosomal DNA per tumour has loss of heterozygosity (LOH). Unsurprisingly, TP53 has the highest LOH recurrence at 67% of samples, and many of the other LOH peaks are also explained by well-known tumour-suppressor genes (TSGs). However, several clear LOH peaks are observed that cannot easily be explained by known TSG selection, such as one on 8p (57% of samples). LOH at 8p has previously been linked to lipid metabolism and drug responses<sup>18</sup>, although the involvement of individual genes has not been established.

There are remarkable differences in the LOH between cancer types (Supplementary Fig. 1). For instance, we observed LOH events on the 3p arm in 90% of kidney samples<sup>19</sup> and LOH of the complete chromosome 10 in 72% of CNS tumours (predominantly glioblastoma multiforme<sup>20</sup>). Furthermore, the mechanism for LOH in *TP53* is highly specific to tumour type, with ovarian cancers exhibiting LOH of the

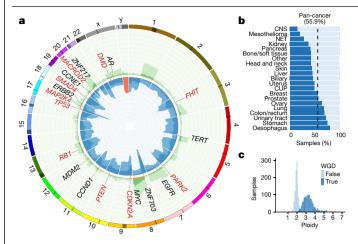


Fig. 2 | Copy number landscape of metastatic cancer. a, Proportion of a concers. asamples with amplification and deletion events by genomic position pancancer. The inner ring shows the percentage of tumours with homozygous deletion (orange), LOH and significant loss (copy number < 0.6× sample ploidy; dark blue) and near copy neutral LOH (light blue). Outer ring shows percentage of tumours with high level amplification (>3× sample ploidy; orange), moderate amplification (>2× sample ploidy; dark green) and low level amplification (>1.4× amplification; light green). The scale on both rings is 0-100% and inverted for the inner ring. The most frequently observed high-level gene amplifications (black text) and homozygous deletions (red text) are shown. b, Proportion of tumours with a WGD event (dark blue), grouped by tumour type. c, Sample ploidy distribution over the complete cohort for samples with and without

full chromosome 17 in 75% of samples, whereas in prostate cancer (also 70% LOH for TP53) this is nearly always caused by highly focal deletions.

Unlike LOH events, homozygous deletions are nearly always restricted to small chromosomal regions. Not a single example was found in which a complete autosomal arm was homozygously deleted. Homozygous deletions of genes are also surprisingly rare: we found only a mean of 2.0 instances per tumour in which one or several consecutive genes are fully or partially homozygously deleted. In 46% of these events, a putative TSG was deleted. Loss of chromosome Y is a special case and is deleted in 36% of all male tumour genomes but varies strongly between tumour types, from 5% deleted in CNS tumours to 68% deleted in biliary tumours (Extended Data Fig. 7).

An extreme form of copy number change can be caused by wholegenome duplication (WGD). We found WGD events in 56% of all samples ranging from 15% in CNS to 80% in oesophageal tumours (Fig. 2). This is  $much \, higher \, than \, previously \, reported \, for \, primary \, tumours \, (25-37\%)^{21,22}$ and from panel-based sequencing analyses of advanced tumours (30%)<sup>23</sup>.

#### Significantly mutated genes

Analyses for significantly mutated genes using strict significance cut-off values (q < 0.01) reproduced previous results on cancer drivers<sup>24</sup>, and identified a few novel genes that are potentially related to metastatic cancer (Extended Data Fig. 8, Supplementary Table 3). In the pan-cancer analyses, we identified MLK4 (also known as MAP3K21;  $q = 2 \times 10^{-4}$ )—a mixed lineage kinase that regulates the JNK, P38 and ERK signalling pathways and has been reported to inhibit tumorigenesis in colorectal cancer<sup>25</sup>. In addition, in our tumour type-specific analyses, we identified a metastatic breast cancer-specific significantly mutated gene-ZFPM1 (also known as *FOG1*;  $q = 8 \times 10^{-5}$ ), a zinc-finger transcription factor protein without clear links to cancer. Our cohort also lends support to previous findings for significantly mutated genes that are currently not included in the COSMIC Cancer Gene Census<sup>26</sup>. In particular, eight significantly mutated putative TSGs found previously in an independent dataset<sup>24</sup> were also found in our analyses, including GPS2 (pan-cancer, breast), SOX9 (pan-cancer, colorectal), TGIF1 (pan-cancer, colorectal), ZFP36L1 (pan-cancer, urinary tract) and ZFP36L2 (pan-cancer, colorectal), HLA-B (lymphoid), MGA (pan-cancer), KMT2B (skin) and RARG (urinary tract).

We also searched for genes that were significantly amplified or deleted (Supplementary Table 4). CDKN2A and PTEN were the most significantly deleted genes overall, but many of the top genes involved common fragile sites, particularly FHIT and DMD, which were deleted in 5% and 4% of samples, respectively. The role of common fragile sites in tumorigenesis is unclear and aberrations that affect these genes are frequently treated as passenger mutations that reflect localized genomic instability<sup>27</sup>. In CTNNB1, we identified a recurrent in-frame deletion of the complete exon 3 in 12 samples, 9 of which are colorectal cancers. Notably, these deletions were homozygous but thought to be activating as CTNNB1 normally acts as an oncogene in the WNT and β-catenin pathway and none of these nine colorectal samples had any APC driver mutations. We also identified several significantly deleted genes not previously reported, including MLLT4 (n = 13) and PARD3 (n = 9).

Unlike homozygous deletions, amplification peaks tend to be broad and often encompass large numbers of genes, making identification of the amplification target challenging. However, SOX4 (6p22.3) stands out as a significantly amplified single gene peak (26 amplifications) and is highly enriched in urinary tract cancers (19% of samples highly amplified). SOX4 is known to be overexpressed in prostate, hepatocellular, lung, bladder and medulloblastoma cancers with poor prognostic features and advanced disease status and is a modulator of the PI3K and Akt signalling pathway<sup>28</sup>.

Also notable was a broad amplification peak of 10 genes around ZMIZ1 at 10q22.3 (n = 32), which has not previously been reported. ZMIZ1 is a transcriptional coactivator of the protein inhibitor of activated STAT (PIAS)-like family and is a direct and selective cofactor of NOTCH1 in the development of T cells and leukaemia<sup>29</sup>. CDX2, previously identified as an amplified lineage-survival oncogene in colorectal cancer<sup>30</sup>, is also highly amplified in our cohort with 20 out of 22 amplified samples found in colorectal cancer, representing 5.4% of all colorectal samples.

#### **Driver mutation catalogue**

We created a comprehensive catalogue of mutations in known (COSMIC curated genes<sup>31</sup>) and newly discovered (ref. <sup>24</sup> and this study) cancer genes across all samples and variant classes, similar to that previously described for primary tumours<sup>32</sup> (N. Lopez, personal communication). We used a prioritization scheme to give a likelihood score for each mutation being a potential driver event. By taking into account the proportion of SNVs and indels estimated to be passengers using the dNdScv R package, we found 13,384 somatic candidate driver events among the 20,071 identified mutations in the combined gene panel (Supplementary Table 5), together with 189 germline predisposition variants (Supplementary Table 6). The somatic candidate drivers include 7,400 coding mutations, 615 non-coding point-mutation drivers, 2,700 homozygous deletions (25% of which are in common fragile sites), 2,392 focal amplifications and 276 fusion events. For non-coding variants, only essential splice sites and promoter mutations in TERT were included in the study owing to the current lack of robust evidence for other recurrent oncogenic non-coding mutations<sup>33</sup>. A total of 257 variants were found at 5 known recurrent variant hotspots9 and included in the candidate driver catalogue.

For the cohort as a whole, 55% of point mutations in the gene panel candidate driver catalogue were predicted to be genuine driver events, using our prioritization scheme (Methods). To facilitate the analysis of variants of unknown significance at a per-patient level, we calculated a sample-specific likelihood score for each point mutation being a driver event by taking into account the mutational burden of the sample, the biallelic inactivation status for TSGs, and hotspot positions for oncogenes. Predictions of pathogenic variant overlap with known

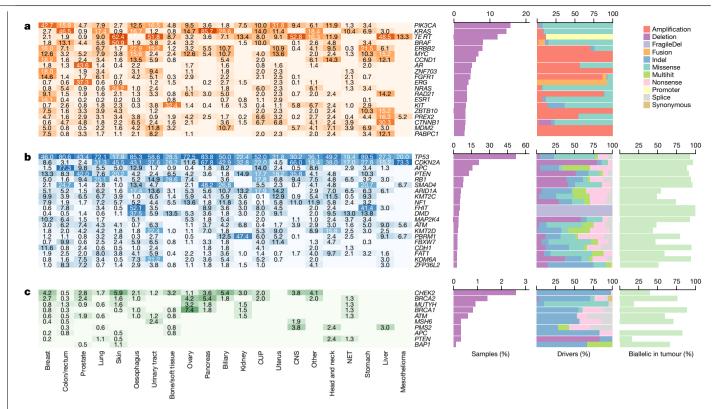


Fig. 3 | The most prevalent driver genes in metastatic cancer. a-c, The most prevalent somatically mutated oncogenes (a), TSGs (b) and germline predisposition variants (c). From left to right, the heat map shows the percentage of samples in each cancer type that are found to have each gene mutated; absolute bar chart shows the pan-cancer percentage of samples with

the given gene mutated; relative bar chart shows the breakdown by type of alteration. For TSGs (b), the final bar chart shows the percentage of samples with a driver in which the gene is biallelically inactivated, and for germline predisposition variants (c), the final bar chart shows the percentage of samples with loss of wild type in the tumour.

biology-for example, clustering of benign missense variants in the 3' half of the APC gene (Supplementary Fig. 2)—fits with the absence of FAP-causing germline variants in this part of the gene<sup>34</sup>.

Overall, the catalogue is similar to previous inventories of cancer drivers, with TP53 (52% of samples), CDKN2A (21%), PIK3CA (16%), APC (15%), KRAS (15%), PTEN (13%) and TERT (12%) identified as the most commonly mutated genes, which together make up 26% of all the candidate driver mutations in the catalogue (Fig. 3). However, all of the ten most frequently mutated genes in our catalogue were reported at a higher rate than for primary cancers<sup>35</sup>, which may reflect the more advanced disease state. AR and ESR1 in particular are more prevalent, with putative driver mutations in 44% of prostate and 16% of breast cancers, respectively. Both genes are linked to resistance to hormonal therapy, a common treatment for these tumour types, and have been previously reported as enriched in advanced metastatic cancer<sup>9</sup> but are identified at higher rates in this study.

At the per-patient level, the mean number of total candidate driver events per patient was 5.7, with the highest rate in urinary tract tumours (mean value of 8.0) and the lowest in NETs (mean of 2.8) (Fig. 4). Oesophageal and stomach tumours also had increased driver counts, largely owing to a much higher rate of deletions in common fragile site genes (mean of 1.6 for both stomach and oesophageal tumours) compared with other cancer types (pan-cancer mean of 0.3). Fragile sites aside, the differential rates of drivers between cancer types in each variant class do correlate with the relative mutational load (Extended Data Fig. 4), with the exception of skin cancers, which have a lower than expected number of SNV drivers.

In 98.6% of all samples, at least one somatic candidate driver mutation or germline predisposition variant was found. Of the 34 samples with no identified driver, 18 were NETs of the small intestine (representing 49% of all patients of this subtype). This probably indicates that small intestine NETs have a distinct set of yet drivers that are not captured in any of the cancer gene resources used and are also not prevalent enough in our relatively small NET cohort to be detected as significant. Alternatively, NETs could be mainly driven by epigenetic mechanisms that are not detected by WGS<sup>36</sup>.

The number of amplified driver genes varied significantly between cancer types (Extended Data Fig. 7), with highly increased rates per sample in breast cancer (mean of 2.1), oesophageal cancer (mean of 1.8), urinary tract and stomach cancers (both mean of 1.7), nearly no amplification drivers in kidney cancer (mean of 0.1), and none in the

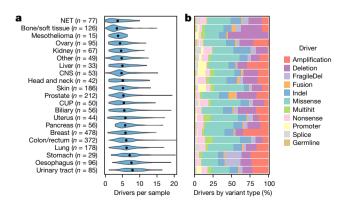
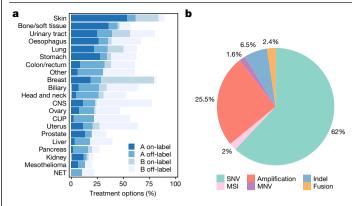


Fig. 4 | Number of drivers and types of mutation per sample by tumour type. a, Violin plot showing the distribution of the number of drivers per sample grouped by tumour type (number of patients per tumour type is provided). Black dots indicate the mean values for each tumour type. **b**, Relative bar chart showing the breakdown per cancer type of the type of alteration.



**Fig. 5** | **Clinical associations and actionability. a**, Percentage of samples in each cancer type with a putative candidate actionable mutation based on data in the CGI, CIViC and OncoKB databases. Level A represents presence of biomarkers with either an approved therapy or guidelines, and level B represents biomarkers with strong biological evidence or clinical trials that indicate that they are actionable. On-label indicates treatment registered by federal authorities for that tumour type, whereas off-label indicates a registration for other tumour types. **b**, Break down of the actionable variants by variant type.

mesothelioma cohort. In tumour types with high rates of amplifications, these amplifications are generally found across a broad spectrum of oncogenes, which suggests that there are mutagenic processes active in these tissues that favour amplifications, rather than tissue-specific selection of individual driver genes. AR and EGFR are notable exceptions, with highly selective amplifications in prostate cancer, and in CNS and lung cancers, respectively, in line with previous reports  $^{20,37,38}$ . Notably, we also found twofold more amplification drivers in samples with WGD events despite amplifications being defined as relative to the average genome ploidy.

The 189 germline variants identified in 29 cancer predisposition genes (present in 7.9% of the cohort) consisted of 8 deletions and 181 point mutations (Fig. 3c, Supplementary Table 6). The top five affected genes (containing nearly 80% of variants) were the well-known germline drivers *CHEK2, BRCA2, MUTYH, BRCA1* and *ATM*. The corresponding wild-type alleles were found to be lost in the tumour sample in more than half of the cases, either by LOH or somatic point mutation, indicating a high penetrance for these variants, particularly in *BRCA1* (89% of cases), *APC* (83%) and *BRCA2* (79%).

The 276 fusions consisted of 168 in-frame coding fusions, 90 *cis*-activating fusions that involve repositioning of regulatory elements in 5′ genic regions, and 18 in-frame intragenic deletions in which one or more exons was deleted (Supplementary Table 7). ERG (n=88), BRAF (n=17), ERBB4 (n=16), ERBB4

We found that 71% of somatic driver point mutations in oncogenes occur at or within five nucleotides already known to pathogenic mutational hotspots. In the six most prevalent oncogenes (KRAS, PIK3CA, BRAF, NRAS, TERT and ESRI), the rate was 97% (Extended Data Fig. 9). Furthermore, in many of the key oncogenes, we document several likely activating but non-canonical variants near known mutational hotspots, particularly in-frame indels. Despite in-frame indels being exceptionally rare overall (mean of 1.7 per tumour), we found an excess in known oncogenes including PIK3CA (n=18), KIT (n=17), ERBB2 (n=10) and BRAF (n=8) frequently occurring at or near known hotspots (Extended Data Fig. 9). In FOXAI, we identified ten in-frame indels that are highly enriched in prostate cancer (seven out of ten cases) and clustered at two locations that were not previously associated with pathogenic mutations  $^{42}$ .

For TSGs, our results strongly support the Knudson two-hit hypothesis<sup>43</sup>, with 80% of all TSG drivers found to have biallelic inactivation by genetic alterations (Fig. 3), homozygous deletion (32%), multiple somatic point mutations (7%), or a point mutation in combination with LOH (41%). This rate is, to our knowledge, the highest observed in any large-scale WGS cancer study. For many key TSGs, the biallelic inactivation rate is almost 100%–*TPS3* (93%), *CDKN2A* (97%), *RB1* (94%), *PTEN* (92%) and *SMAD4* (96%)—which suggests that biallelic genetic inactivation of these genes is a strong requirement for metastatic cancer. Other prominent TSGs, however, have lower biallelic inactivation rates, including *ARID1A* (55%), *KMT2C* (49%) and *ATM* (49%). For these cases, the other allele may also be inactivated by non-mutational epigenetic mechanisms, or tumorigenesis may be driven via a haploinsufficiency mechanism.

We examined the pairwise co-occurrence of driver gene mutations per cancer type and found ten combinations of genes that were significantly mutually exclusively mutated, and ten combinations of genes that were significantly concurrently mutated (Extended Data Fig. 10). Although most of these relationships are well established, in breast cancer, we found new positive relationship for GATA3-VMP1 ( $q=6\times10^{-5}$ ) and FOXA1-PIK3CA ( $q=3\times10^{-3}$ ), and negative relationships for ESRI-TP53 ( $q=9\times10^{-4}$ ) and GATA3-TP53 ( $q=5\times10^{-5}$ ). These findings will need further validation and experimental follow-up to understand the underlying biology.

#### **Clonality of variants**

To obtain insight into ongoing tumour evolution dynamics, we examined the clonality of all variants. Notably, only 6.6% of all SNVs, MNVs and indels across the cohort and just 3.7% of the point-mutation drivers were found to be subclonal (Extended Data Fig. 11). The low proportion of samples with subclonal variants could be partially due to the detection limits of the sequencing approach (sequencing depth, bioinformatic analysis settings), particularly for low purity samples. However, even for samples with more than 80% purity, the total proportion of subclonal variants only reaches 10.6% (Extended Data Fig. 11). Furthermore, sensitized detection of variants at hotspot positions in cancer genes showed that our analysis pipeline detected over 96% of variants with allele frequencies above 3%. Although the cohort contains some samples with high fractions of subclonal variants, overall the metastatic tumour samples are relatively homogeneous without the presence of multiple diverged major subclones. Low intratumour heterogeneity may be in part attributed to the fact that nearly all biopsies were obtained by a core needle biopsy, which results in highly localized sampling, but is nevertheless much lower than previous observations in primary cancers<sup>12</sup>.

In the 117 patients with independently collected repeat biopsies from the same patient (Supplementary Table 8), we found 11% of all SNVs to be subclonal. Although 71% of clonal variants were shared between biopsies, only 29% of the subclonal variants were shared. We cannot exclude the presence of larger amounts of lower frequency subclonal variants, and our results suggest a model in which individual metastatic lesions are dominated by a single clone at any one point in time and that more limited tumour evolution and subclonal selection takes places after distant metastatic seeding. This contrasts with observations in primary tumours, in which larger degrees of subclonality and several major subclones are more frequently observed 12,44, but supports other recent studies that demonstrate minimal driver gene heterogeneity in metastases 6,45.

#### **Clinical associations**

We analysed opportunities for biomarker-based treatment for all patients by mapping driver events to clinical annotation databases (CGI<sup>41</sup>, CIViC<sup>39</sup> and OncoKB<sup>40</sup>). In 1,480 patients (62%), at least one predicted candidate 'actionable' event was identified (as defined in the Methods, Supplementary Table 9), in line with results from primary

tumours<sup>32</sup>. Half of the patients with a predicted candidate actionable event (31% of total) contained a biomarker with a predicted sensitivity to a drug at level A (approved anti-cancer drugs) and lacked any known resistance biomarkers for the same drug (Fig. 5a). In 18% of patients, the suggested therapy was a registered indication, whereas in 13% of cases it was outside the labelled indication. In a related pilot study with implementation in 215 treated patients, we showed that such treatment with anticancer drugs outside of their approved label can result in overall clinical benefits<sup>46</sup>. In a further 31% of patients, a level B (experimental therapy) biomarker was identified. The predicted actionable events spanned all variant classes including 1,815 SNVs, 48 MNVs, 190 indels, 745 copy number alterations, 69 fusion genes and 60 patients with microsatellite instability (Fig. 5b).

Tumour mutation burden (TMB) is an important emerging biomarker for responses to immune checkpoint inhibitor therapy as it is a proxy for the amount of neo-antigens in the tumour cells. In two large phase 3 trials of patients with non-small-cell lung cancer, both progression-free survival and overall survival are significantly improved with first line immunotherapy as compared with chemotherapy for patients whose tumours have a TMB of greater than 10 mutations per megabase 47,48.

Although various clinical studies based on this parameter are currently emerging, TMB was not yet included in the above actionability analysis. However, when applying this cut-off to all samples in our cohort, 18% of patients would qualify, varying from 0% for patients with mesothelioma, liver and ovarian cancers to more than 50% for patients with lung and skin cancers (Extended Data Fig. 4b).

#### Data availability and resource access

The Hartwig Medical cohort described here is, to our knowledge, the  $largest\,metastatic\,whole-genome\,cancer\,resource, and\,based\,on\,a\,broad$ patient consent was specifically developed as a community resource for international academic cancer research. Somatic variants and basic clinical data (tumour type, gender, age) are publicly available and can be explored at the patient, cohort and gene level through a graphical interface (database.hartwigmedicalfoundation.nl) originally developed by the International Cancer Genome Consortium<sup>49</sup>. Patient-level genome-wide germline and somatic data (raw BAM files and annotated variant call data) are considered privacy sensitive and available through an access-controlled mechanism (see www.hartwigmedicalfoundation. nl/en for details).

The cohort is still expanding, with data from 4,000 patients already available, and includes data that go beyond the basic clinical and genomic data analysed in this paper such as post-biopsy treatments and responses, and previous treatment information.

#### **Discussion**

Genomic testing of tumours faces numerous challenges in meeting clinical needs, including the interpretation of variants of unknown significance, the steadily expanding universe of actionable genes-often with an increasingly small fraction of patients affected—and the development of advanced genome-derived biomarkers such as tumour mutational load, DNA repair status and mutational signatures. Our results demonstrate that WGS analyses of metastatic cancer can provide novel and relevant insights and are instrumental in addressing some of the key challenges of precision medicine in cancer.

 $First, our systematic and large-scale pan-cancer analyses \, on \, metastatic$ cancer tissue allowed for the identification of several cancer drivers and mutation hotspots. Second, the driver catalogue analyses can be used to mitigate the problem of variants of unknown significance interpretation<sup>32</sup> both by leveraging previously identified pathogenic mutations (accounting for more than two-thirds of oncogenic point-mutation drivers) and by careful analysis of the biallelic inactivation of putative TSGs that accounts for over 80% of TSG drivers in metastatic cancer.

Third, we demonstrate the importance of accounting for all types of variant, including large-scale genomic rearrangements (via fusions and copy number alteration events), which account for more than half of all drivers, but also activating MNVs and indels that we have shown are commonly found in many key oncogenes. Fourth, we have shown that using WGS, even with very strict variant calling criteria, we could find candidate driver variants in more than 98% of all metastatic tumours, including predicted putatively actionable events in a clinical and experimental setting for up to 62% of patients.

Although we did not find metastatic tumour genomes to be fundamentally different from primary tumours in terms of the mutational landscape or genes that drive advanced tumorigenesis, we described characteristics that could contribute to responsiveness to therapy or resistance in individual patients. In particular, we showed that WGD events are a more pervasive element of tumorigenesis than previously understood, affecting over half of all metastatic cancers. We also found metastatic lesions to be less heterogeneous than reported for primary tumours, although the limited sequencing depth does not allow conclusions to be made about low-frequency subclonal variants.

The cohort described here provides a valuable complementary resource to whole-sequence-based data of primary tumours such as the PCAWG project in advancing fundamental and translational cancer research. Although it was established as a pan-cancer resource, several of the tumour type-specific cohorts are very large in their own rights. Already two of these cohorts (prostate<sup>50</sup> and breast<sup>51</sup>) have been analysed in more detail, providing enhanced cancer subtype stratification and revealing characteristic genomic differences between primary and metastatic tumours. As the Hartwig Medical cohort includes a mix of treatmentnaive metastatic patients and patients who have undergone (extensive) previous systemic treatments, it provides unique opportunities to study responses and resistance to treatments and discover predictive biomarkers, as these data are available for discovery and validation studies.

#### **Online content**

Any methods, additional references, Nature Research reporting summaries, source data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1689-y.

- The Cancer Genome Atlas Research Network et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. 45, 1113-1120 (2013).
- The International Cancer Genome Consortium. International network of cancer genome projects, Nature 464, 993-998 (2010).
- Gröbner, S. N. et al. The landscape of genomic alterations across childhood cancers. Nature 555, 321-327 (2018).
- Ma, X. et al. Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. Nature 555, 371-376 (2018).
- Hyman, D. M., Taylor, B. S. & Baselga, J. Implementing genome-driven oncology, Cell 168. 584-599 (2017).
- Yates, L. R. et al. Genomic evolution of breast cancer metastasis and relapse. Cancer Cell 32. 169–184 (2017).
- Naxerova, K. et al. Origins of lymphatic and distant metastases in human colorectal cancer Science 357, 55-60 (2017)
- Gundem, G. et al. The evolutionary history of lethal metastatic prostate cancer. Nature 520 353-357 (2015)
- Zehir, A. et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nat. Med. 23, 703-713 (2017).
- 10. Robinson, D. R. et al. Integrative clinical genomics of metastatic cancer. Nature 548, 297-303 (2017).
- Klein, C. A. Selection and adaptation during metastatic cancer progression. Nature 501, 365-372 (2013). McGranahan, N. & Swanton, C. Clonal heterogeneity and tumor evolution: past, present,
- and the future. Cell 168, 613-628 (2017). Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. Nature 500, 415-421 (2013).
- Campbell, P. J., Getz, G., Stuart, J. M., Korbel, J. O. & Stein, L. D. Pan-cancer analysis of whole genomes. Preprint at https://www.bioRxiv.org/content/10.1101/162784v1
- Kucab, J. E. et al. A compendium of mutational signatures of environmental agents. Cell 177, 821-836 (2019)

- Gryfe, R, et al. Tumor microsatellite instability and clinical outcome in young patients with colorectal cancer. N. Engl. J. Med. 342, 69-77 (2000).
- 17. Wedge, D. C. et al. Sequencing of prostate cancers identifies new cancer genes, routes of progression and drug targets. Nat. Genet. 50, 682-692 (2018).
- Cai, Y. et al. Loss of chromosome 8p governs tumor progression and drug response by 18. altering lipid metabolism. Cancer Cell 29, 751-766 (2016).
- Sato, Y. et al. Integrated molecular analysis of clear-cell renal cell carcinoma. Nat. Genet. 45, 860-867 (2013)
- 20. Brennan, C. W. et al. The somatic genomic landscape of glioblastoma. Cell 155, 462-477
- Zack, T. I. et al. Pan-cancer patterns of somatic copy number alteration. Nat. Genet. 45, 1134-1140 (2013).
- 22. Carter, S. L. et al. Absolute quantification of somatic DNA alterations in human cancer. Nat. Biotechnol. 30, 413-421 (2012).
- Bielski, C. M. et al. Genome doubling shapes the evolution and prognosis of advanced 23 cancers. Nat. Genet. 50, 1189-1195 (2018).
- 24. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. Cell 171, 1029-1041.e21 (2017).
- 25 Marusiak, A. A. et al. Recurrent MLK4 loss-of-function mutations suppress JNK signaling to promote colon tumorigenesis. Cancer Res. 76, 724-735 (2016).
- 26 Forbes, S. A. et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. 45 (D1), D777-D783 (2017).
- Glover, T. W., Wilson, T. E. & Arlt, M. F. Fragile sites in cancer: more than meets the eye. 27
- Nat. Rev. Cancer 17, 489-501 (2017) 28 Mehta, G. A. et al. Amplification of SOX4 promotes PI3K/Akt signaling in human breast cancer. Breast Cancer Res. Treat. 162, 439-450 (2017).
- Pinnell, N. et al. The PIAS-like coactivator Zmiz1 is a direct and selective cofactor of Notch1 in T cell development and leukemia. Immunity 43, 870-883 (2015)
- 30. Salari, K. et al. CDX2 is an amplified lineage-survival oncogene in colorectal cancer. Proc. Natl Acad. Sci. USA 109, E3196-E3205 (2012).
- Futreal, P. A. et al. A census of human cancer genes. Nat. Rev. Cancer 4, 177-183 (2004).
- Sabarinathan, R. et al. The whole-genome panorama of cancer drivers. Preprint at https:// www.bioRxiv.org/content/10.1101/190330v2 (2017).
- Cuykendall, T. N., Rubin, M. A. & Khurana, E. Non-coding genetic variation in cancer. Current Opinion in Systems Biology 1, 9-15 (2017).
- Friedl, W. et al. Can APC mutation analysis contribute to therapeutic decisions in familial adenomatous polyposis? Experience from 680 FAP families. Gut 48, 515-521 (2001).
- 35. Bailey, M. H. et al. Comprehensive characterization of cancer driver genes and mutations. Cell 173, 371-385 (2018).
- Cives, M., Simone, V., Rizzo, F. M. & Silvestris, F. NETs: organ-related epigenetic derangements and potential clinical applications. Oncotarget 7, 57414-57429 (2016).
- 37. Viswanathan, S. R. et al. Structural alterations driving castration-resistant prostate cancer revealed by linked-read genome sequencing, Cell 174, 433-447 (2018).
- The Cancer Genome Atlas Research Network, Comprehensive genomic characterization of squamous cell lung cancers. Nature 489, 519-525 (2012).

- Griffith, M. et al. CIVIC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat. Genet. 49, 170-174 (2017).
- Chakravarty, D. et al. OncoKB: a precision oncology knowledge base. JCO Precis. Oncol. 40. https://doi.org/10.1200/PO.17.00011 (2017).
- Tamborero, D. et al. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. Genome Med. 10, 25 (2018).
- Yang, Y. A. & Yu, J. Current perspectives on FOXA1 regulation of androgen receptor signaling and prostate cancer. Genes Dis. 2, 144–151 (2015).
- Knudson, A. G. Jr Mutation and cancer: statistical study of retinoblastoma. Proc. Natl Acad. Sci. USA 68, 820-823 (1971).
- Andor, N. et al. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. Nat. Med. 22, 105-113 (2016).
- Reiter, J. G. et al. Minimal functional driver gene heterogeneity among untreated metastases. Science 361, 1033-1037 (2018).
- van der Velden, D. L. et al. The Drug Rediscovery protocol facilitates the expanded use of existing anticancer drugs, Nature 574, 127-131 (2019).
- Hellmann, M. D. et al. Nivolumab plus ipilimumab in lung cancer with a high tumor mutational burden, N. Engl. J. Med. 378, 2093-2104 (2018)
- Carbone, D. P. et al. First-line nivolumab in stage IV or recurrent non-small-cell lung cancer, N. Engl. J. Med. 376, 2415-2426 (2017).
- Zhang, J. et al. The International Cancer Genome Consortium Data Portal. Nat. Biotechnol. 37, 367-369 (2019).
- van Dessel, L. F. et al. The genomic landscape of metastatic castration-resistant prostate cancers using whole genome sequencing reveals multiple distinct genotypes with potential clinical impact. Preprint at https://www.bioRxiv.org/content/10.1101/546051v1 (2019)
- Angus, L. et al. Genomic landscape of metastatic breast cancer and its clinical implications. Nat. Genet. 51, 1450-1458 (2019).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution

and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use. you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019

#### Methods

A detailed description of methods and validations is available as Supplementary Information. No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

#### Sample collection

Patients with advanced cancer not curable by local treatment options and being candidates for any type of systemic treatment and any line of treatment were included as part of the CPCT-02 (NCT01855477) and DRUP (NCT02925234) clinical studies, which were approved by the medical ethical committees (METC) of the University Medical Center Utrecht and the Netherlands Cancer Institute, respectively, A total of 41 academic, teaching and general hospitals across The Netherlands participated in these studies and collected material and clinical data by standardized protocols<sup>52</sup>. Patients have given explicit consent for whole-genome sequencing and data sharing for cancer research purposes. Core needle biopsies were sampled from the metastatic lesion, or when considered not feasible or not safe, from the primary tumour site and frozen in liquid nitrogen. A single 6-μm section was collected for haematoxylin and eosin (H&E) staining and estimation of tumour cellularity by an experienced pathologist and 25 sections of 20-µm were collected in a tube for DNA isolation. In parallel, a tube of blood was collected. Leftover material (biopsy, DNA) was stored in biobanks associated with the studies at the University Medical Center Utrecht and the Netherlands Cancer Institute.

#### Whole-genome sequencing and variant calling

DNA was isolated from biopsies (>30% tumour cellularity) and blood according to the supplier's protocols (Qiagen) using the DSP DNA Midi kit for blood and QIAsymphony DSP DNA Mini kit for tissue. A total of 50-200 ng of DNA (sheared to average fragment length of 450nt) was used as input for TruSeq Nano LT library preparation (Illumina). Barcoded libraries were sequenced as pools on HiSeqX generating 2×150 read pairs using standard settings (Illumina). BCL output was converted using bcl2fastq tool (Illumina, v.2.17 to v.2.20) using default parameters. Reads were mapped to the reference genome GRCH37 using BWA-mem v.0.7.5a<sup>53</sup>, duplicates were marked for filtering and INDELs were realigned using GATK v.3.4.46 IndelRealigner<sup>54</sup>. GATK HaplotypeCaller v.3.4.46<sup>55</sup> was run to call germline variants in the reference sample. For somatic SNV and indel variant calling, GATK BQSR<sup>56</sup> was applied to recalibrate base qualities. SNV and indel somatic variants were called using Strelka v.1.0.14<sup>57</sup> with optimized settings and post-calling filtering. Structural Variants were called using Manta (v.1.0.3)<sup>58</sup> with default parameters followed by additional filtering to improve precision using an internally built tool (Breakpoint-Inspector v.1.5). To assess the effect of sequencing depth on variant calling sensitivity, we downsampled the BAMS of 10 samples at random by 50% and reran the identical somatic variant calling pipeline.

#### Purity, ploidy and copy number calling

Copy number calling and determination of sample purity were performed using PURPLE (PURity & PLoidy Estimator), which combines B-allele frequency, read depth and structural variants to estimate the purity of a tumour sample and determine the copy number and minor allele ploidy for every base in the genome. The purity and ploidy estimates and copy number profile obtained from PURPLE were validated on in silico simulated tumour purities, by DNA fluorescence in situ hybridization (FISH) and by comparison with an alternative tool (ASCAT<sup>59</sup>). ASCAT was run on GC-corrected data using default parameters except for gamma, which was set to 1 (which is recommended for massively parallel sequencing data). We implement a simple heuristic that determines if a WGD event has occurred: major allele ploidy > 1.5 on at least 50% of at least 11 autosomes as the number of duplicated autosomes

per sample (that is, the number of autosomes which satisfy the above rule) follows a bimodal distribution with 95% of samples have either  $\leq$ 6 or  $\geq$ 15 autosomes duplicated.

#### Sample selection for downstream analyses

Following copy number calling, samples were filtered out based on absence of somatic variants, purity <20%, and GC biases, yielding a high-quality dataset of 2,520 samples. Where multiple biopsies exist for a single patient, the highest purity sample was used for downstream analyses (resulting in 2,399 samples).

#### Mutational signature analysis

Mutational signatures were determined by fitting SNV counts per 96 tri-nucleotide context to the 30 COSMIC signatures  $^{26}$  using the mutational Patterns package  $^{60}$ . Residuals were calculated as the sum of the absolute difference between observed and fitted across the 96 buckets. Signatures with <5% overall contribution to a sample or absolute fitted mutational load <300 variants were excluded from the summary plot.

#### Germline predisposition variant calling

We searched for pathogenic germline variants (SNVs, indels and copy number alterations) in a broad list of 152 germline predisposition genes previously curated<sup>61</sup>, using GATK HaplotypeCaller<sup>55</sup> output from each sample. For each variant identified, we assessed the genotype in the germline (HET or HOM), whether there was a second somatic hit in the tumour, and whether the wild type or the variant itself was lost by a copy number alteration. We observed that for the variants in many of the 152 predisposition genes that a loss of wild type in the tumour via LOH was lower than the average rate of LOH across the cohort and that fewer than 5% of observed variants had a second somatic hit in the same gene. Moreover, in many of these genes, the ALT variant was lost via LOH as frequently as the wild type, suggesting that a considerable portion of the 566 variants may be passengers. For our downstream analysis and driver catalogue, we therefore restricted our analysis to a more conservative 'high confidence' list including only the 25 cancer related genes in the ACMG secondary findings reporting guidelines (v.2.0)<sup>62</sup>, together with four curated genes (CDKN2A, CHEK2, BAP1 and ATM), selected because these are the only additional genes from the larger list of 152 genes with a significantly increased proportion of called germline variants with loss of wild type in the tumour sample.

#### Clonality and biallelic status of point mutations

The ploidy of each variant is calculated by adjusting the observed VAF by the purity and then multiplying by the local copy number to work out the absolute number of chromatids that contain the variant. We mark a mutation as biallelic (that is, no wild type remaining) if variant ploidy > local copy number – 0.5. For each variant, we also determine a probability that it is subclonal. This is achieved via a two-step process involving fitting the somatic ploidies for each sample into a set of clonal and subclonal peaks and calculating the probability that each individual variant belongs to each peak. Subclonal counts are calculated as the total density of the subclonal peaks for each sample. Subclonal driver counts are calculated as the sum across the driver catalogue of subclonal probability  $\times$  driver likelihood.

#### MSI status determination

To determine the MSI status, we used the method described by the MSIseq tool  $^{63}$  and counted the number of indels per million bases occurring in homopolymers of five or more bases or dinucleotide, trinucleotide and tetranucleotide sequences of repeat count four or more. MSIseq score of >4 were considered MSI.

#### Significantly mutated driver genes

We used Ensembl<sup>64</sup> v.89.37 as a basis for gene definitions and have taken the union of Entrez identifiable genes and protein-coding genes as our

base panel (25,963 genes of which 20,083 genes are protein coding). Pan-cancer and at an individual cancer level we tested the normalized nonsynonymous (dN) to synonymous substitution (dS) rate (that is, dN/dS) using dNdScv<sup>24</sup> against a null hypothesis that dN/dS = 1 for each variant subtype. To identify significantly mutated genes in our cohort, we used a strict significance cut-off value of q < 0.01.

To search for significantly amplified and deleted genes, we first calculated the minimum exonic copy number per gene. For amplifications, we searched for all the genes with high-level amplifications only (defined as minimum exonic copy number >3 × sample ploidy). For deletions, we searched for all the genes in each sample with either full or partial homozygous gene deletions (defined as minimum exonic copy number < 0.5) excluding the Y chromosome. We then searched separately for amplifications and deletions, on a per-chromosome basis, for the most significant focal peaks, using an iterative GISTIC-like peel off method<sup>65</sup>. Most of the deletion peaks resolve clearly to a single target gene, which reflects the fact that homozygous deletions are highly focal, but for amplifications this is not the case and most of our peaks have ten or more candidates. We therefore annotated the peaks, to choose a single putative target gene using an objective set of automated curation rules. Finally, filtering was applied to yield highly significant deletions and amplifications.

Homozygous deletions were also annotated as common fragile sites based on their genomic characteristics, including a strong enrichment in long genes (>500,000 bases) and a high rate (>30%) of deletions between  $20\,kb$  and  $1\,Mb^{27}$ .

#### Somatic driver catalogue construction

We created a catalogue of mutations in known cancer genes in our cohort across all variant types on a per-patient basis. This was done in a similar incremental manner to that previously described  $^{32}$  (N. Lopez, personal communication) in which we first calculated the number of genes with putative driver mutations in a broad panel of known and significantly mutated genes across the full cohort, and then assigned the candidate driver mutations for each gene to individual patients by ranking and prioritizing each of the observed variants. Key points of difference in this study were both the prioritization mechanism used and our choice to ascribe each mutation a probability of being a driver rather than a binary cut-off based on absolute ranking.

The four steps to create the catalogue are as follows. (1) Create a panel of candidate genes for point mutations using significantly mutated genes and known cancer genes using the union of Martincorena significantly mutated genes<sup>24</sup> (filtered to significance of q < 0.01), HMF significantly mutated genes (q < 0.01) at global level or at cancer type level and COSMIC curated genes<sup>26</sup> (v.83). (2) Determine TSG or oncogene status of each significantly mutated gene using a logistic regression classification model (trained using COSMIC annotation). (3) Add mutations from all variant classes to the catalogue when meeting any of the following criteria: (i) all missense and inframe indels for panel oncogenes; (ii) all non-synonymous and essential splice point mutations for TSGs; (iii) all high-level amplifications for significantly amplified target genes and panel oncogenes; (iv) all homozygous deletions for significantly deleted target genes and panel TSGs; (v) all known or promiscuous in-frame gene fusions; and (vi) recurrent TERT promoter mutations. (4) Calculate a per-sample likelihood score (between 0 and 1) for each mutation in the catalogue as a potential driver event, to ensure that only likely pathogenic and excess mutations (based on dN/dS) are used to determine the number of drivers. All putative driver mutation counts reported at a per-cancer type or sample level refer to the sum of driver likelihoods for that cancer type or sample.

#### Clinical associations and actionability analysis

To determine clinical associations and potential actionability of the variants observed in each sample, we compared all variants with

three external clinical annotation databases (OncoKB<sup>40</sup>, CGI<sup>41</sup> and CIViC<sup>39</sup>) that were mapped to a common data model as defined by https://civicdb.org/help/evidence/evidence-levels. Here, we considered only A and B level variants. This classification of potential actionable events can also be mapped to the recently proposed ESMO Scale for Clinical Actionability of molecular Targets (ESCAT)<sup>66</sup> as follows: ESCAT I-A+B (for A on-label) and I-C (for A off-label) and ESCAT II-A+B (for B on-label) and III-A (for B off-label). For each candidate actionable mutation, it was also determined to be either on-label (that is, evidence supports treatment in that specific cancer type) or off-label (evidence exists in another cancer type). To do this, we annotated both the patient cancer types and the database cancer types with relevant DOIDs, using the disease ontology database<sup>67</sup>. For each candidate actionable mutation in each sample, we aggregated all the mapped evidence that was available supporting both on-label and off-label treatments at the A or B evidence level. Treatments that also had evidence supporting resistance based on other biomarkers in the sample at the same or higher evidence level were excluded as non-actionable. Samples classified as MSI in our driver catalogue were also mapped as actionable at level A evidence based on clinical annotation in the OncoKB database. For each sample, we reported the highest level of predicted actionability, ranked first by evidence level and then by on-label vs off-label.

#### **Reporting summary**

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

#### **Data availability**

All data described in this study are freely available for academic use from the Hartwig Medical Foundation through standardized procedures and request forms that can be found at https://www.hartwigmedicalfoundation.nl/en/appyling-for-data/.

Available data include germline and tumour raw sequencing data (BAM files, including non-mapped reads), annotated somatic and germline variants (VCF files with annotated SNV and indels, and pipeline output files for purity and ploidy status as well as copy number alteration and structural variants) and clinical data. Examples of output files can be found at https://resources.hartwigmedicalfoundation.nl. In brief, a data request can be initiated by filling out the standard form in which intended use of the requested data is motivated. First, an advice on scientific feasibility and validity is obtained from experts in the field that is used as input by an independent data access board who also evaluates if the intended use of the data is compatible with the consent given by the patients and if there would be any applicable legal or ethical constraints. Upon formal approval by the data access board, a standard license agreement that does not have any restrictions regarding intellectual property resulting from the data analysis needs to be signed by an official organization representative before access to the data are granted. After approval, access to data is provided under a license model, with the only main restriction that the data can only be used for the research detailed in the original request. Raw data files will be made available through a dedicated download portal with two-factor authentication.

Non-privacy sensitive somatic variants can also be browsed and explored through an open access web-based interface which can be accessed at http://database.hartwigmedicalfoundation.nl/.

#### **Code availability**

All code used is open source and available from third parties or developed by Hartwig Medical Foundation (https://github.com/hartwigmedical/). A full list of tools and versions used including links to the source code is provided in the Supplementary Information.

- Bins, S. et al. Implementation of a multicenter biobanking collaboration for nextgeneration sequencing-based biomarker discovery based on fresh frozen pretreatment tumor tissue biopsies. Oncologist 22, 33-40 (2017).
- Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754–1760 (2009).
- McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303 (2010).
- Poplin, R. et al. Scaling accurate genetic variant discovery to tens of thousands of samples. Preprint at https://www.bioRxiv.org/content/10.1101/201178v2 (2018).
- Van der Auwera, G. A. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinformatics 43, 11.10.1–11.10.33 (2013)
- 57. Saunders, C. T. et al. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28**, 1811–1817 (2012).
- Chen, X. et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* 32, 1220–1222 (2016).
- Van Loo, P. et al. Allele-specific copy number analysis of tumors. Proc. Natl Acad. Sci. USA 107, 16910–16915 (2010).
- Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. Mutational Patterns: comprehensive genome-wide analysis of mutational processes. *Genome Med.* 10, 33 (2018).
- Huang, K.-L. et al. Pathogenic germline variants in 10,389 adult cancers. Cell 173, 355–370 (2018).
- Kalia, S. S. et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. Genet. Med. 19, 249–255 (2017).
- 63. Huang, M. N. et al. MSIseq: software for assessing microsatellite instability from catalogs of somatic mutations. Sci. Rep. 5, 13321 (2015).
- 64. Zerbino, D. R. et al. Ensembl 2018. Nucleic Acids Res. 46 (D1), D754-D761 (2018).
- Mermel, C. H. et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 12, R41 (2011).
- Mateo, J. et al. A framework to rank genomic alterations as targets for cancer precision medicine: the ESMO Scale for Clinical Actionability of molecular Targets (ESCAT). Ann. Oncol. 29, 1895–1902 (2018).
- Kibbe, W. A. et al. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids* Res. 43, D1071–D1078 (2015).

- Black, J. et al. SYD985, a novel duocarmycin-based HER2-targeting antibody-drug conjugate, shows antitumor activity in uterine serous carcinoma with HER2/Neu Expression. Mol. Cancer Ther. 15, 1900–1909 (2016).
- Bond, C. E. et al. RNF43 and ZNRF3 are commonly altered in serrated pathway colorectal tumorigenesis. Oncotarget 7, 70589–70600 (2016).
- Fleming, N. I. et al. SMAD2, SMAD3 and SMAD4 mutations in colorectal cancer. Cancer Res. 73, 725–735 (2013).

Acknowledgements We thank the Hartwig Foundation and Barcode for Life for financial support of clinical studies and WGS analyses. Implementation of the data portal was supported by a grant from KWF Kankerbestrijding (HMF2017-8225, GENONCO). We are particularly grateful to all patients, nurses and medical specialists for their essential contributions that make this study possible, and to H. van Snellenberg (Hartwig Medical Foundation) for operational management. We thank S. Willems, W. de Leng, A. Hoischen and W. Dinjens for support with pathology assessments and mutation validations and J. de Ridder, W. Kloosterman and H. van de Werken for critically reading the manuscript.

Author contributions E.F.S., S.S., E.E.V. and E.C. designed the study. H.J.B., V.C.G.T.-H., C.M.L.v.H., M.L., P.O.W., M.P.L., N.S., E.F.S., S.S. and E.E.V. contributed patient material, M.P.L. and N.S. supervised clinical studies and E.d.B. supervised WGS data generation. P.P., J.B., K.D., S.H., A.V.H., W.O., P.R., C.S. and M.V. structured and analysed data. P.P., J.B. and E.C. wrote the manuscript. All authors provided input for improvement of the manuscript.

Competing interests E.E.V. is a supervisory board member of the Hartwig Medical Foundation.

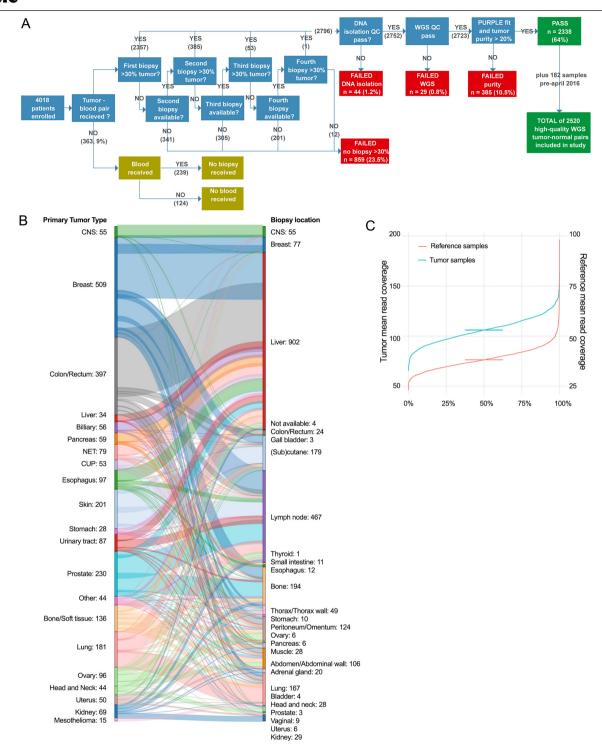
#### Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1689-v

Correspondence and requests for materials should be addressed to E.C.

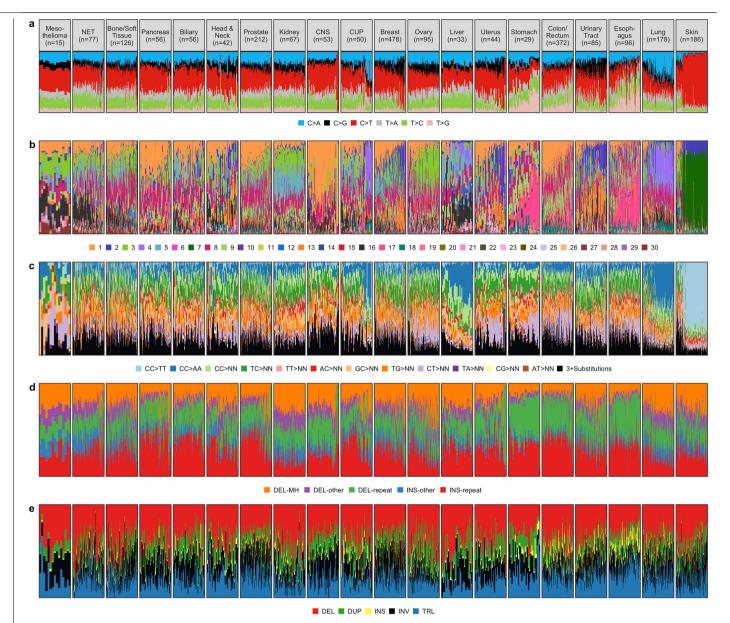
**Peer review information** *Nature* thanks Fran Supek and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints.



**Extended Data Fig. 1**| **Hartwig sample workflow, biopsy locations and sequence coverage. a**, Sample workflow from patient to high-quality WGS data. A total of 4,018 patients were enrolled in the study between April 2016 and April 2018. For 9% of patients, no blood and/or biopsy material was obtained, mostly because conditions of patients prohibited further study participation. Up to four fresh-frozen biopsies were obtained per patient, and were sequentially analysed to identify a biopsy with more than 30% tumour cellularity as determined by routine histology assessment. For 859 patients, no suitable biopsy was obtained, and 2,796 patients were further processed for WGS analysis. In total, 44 and 29 samples failed in either DNA isolation or library preparation and raw WGS data quality control tests, respectively. For a further 385 samples, the WGS data were of good quality, but the determination

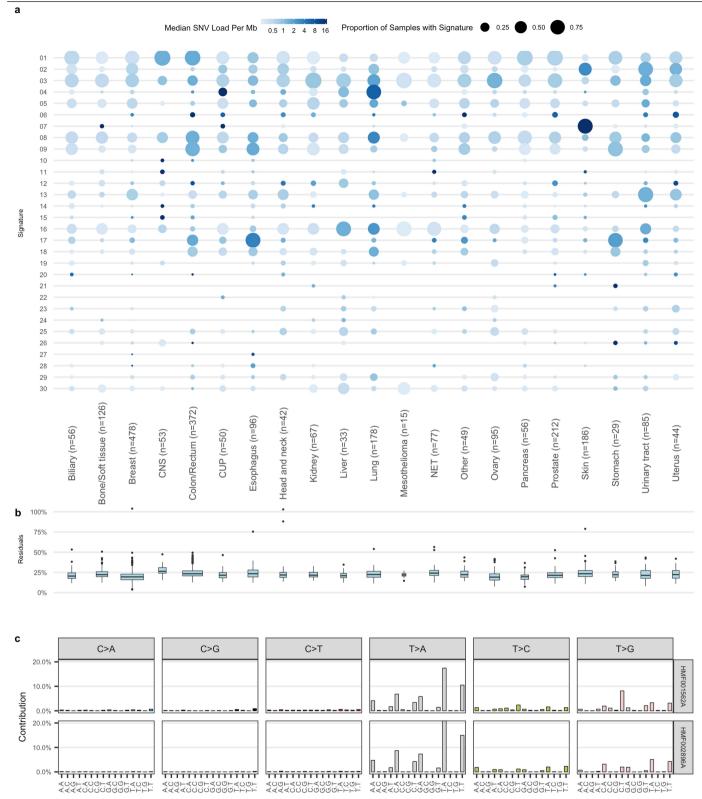
of tumour purity based on WGS data (PURity & PLoidy Estimator; PURPLE) was less than 20%, making reliable and comprehensive somatic variant calling impossible and were therefore excluded. Eventually, 2,338 pairs of tumour and normal tissue samples with high-quality WGS data were obtained, which were supplemented with 182 pairs from pre-April 2016, adding up to 2,520 pairs of tumour and normal samples that were included in this study.  $\bf b$ , Breakdown of cohort by biopsy location. Tumour biopsies were taken from a broad range of locations. Primary tumour type is shown on the left, and the biopsy location on the right.  $\bf c$ , Distribution of sample sequencing depth for tumour and blood reference samples (n = 2,520 independent samples for each category). The median for each is indicated by a horizontal bar.



#### $Extended\,Data\,Fig.\,2\,|\,Mutational\,context\,distribution\,per\,tumour\,type.$

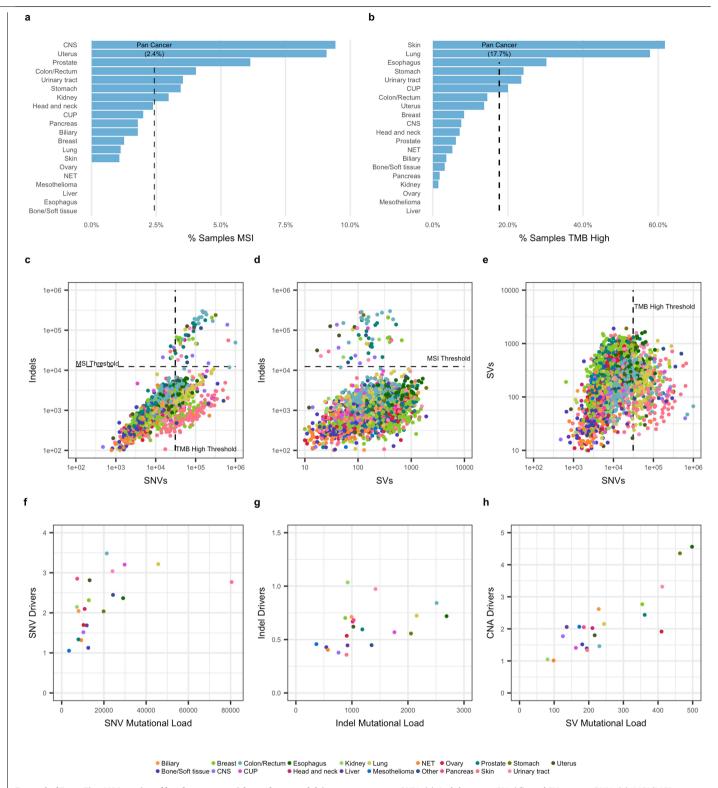
 $\label{eq:a-e} \textbf{a-e}, \textbf{Variant subtype}, \textbf{mutational context or signature per individual sample for } \textbf{each SNV (a)}, \textbf{SNV by COSMIC signature (b)}, \textbf{MNV (c)}, \textbf{indel (d) or SV (e)}. \textbf{Each column chart is ranked within tumour type by mutational load from low to high in that variant class. \textbf{MNVs are classified by the dinucleotide substitution, with 'NN' referring to any dinucleotide combination. SVs are classified by type.}$ 

DEL, deletion (with microhomology (MH), in repeats and other); DUP, tandem duplication; INV, inversion; TRL, translocation; INS, insertion. Highly characteristic known patterns can be discerned, for example the high rates of C>T SNVs, CC>TT MNVs and COSMIC S18 for skin tumours, and high rates of C>A SNVs and COSMIC S4 for lung tumours.



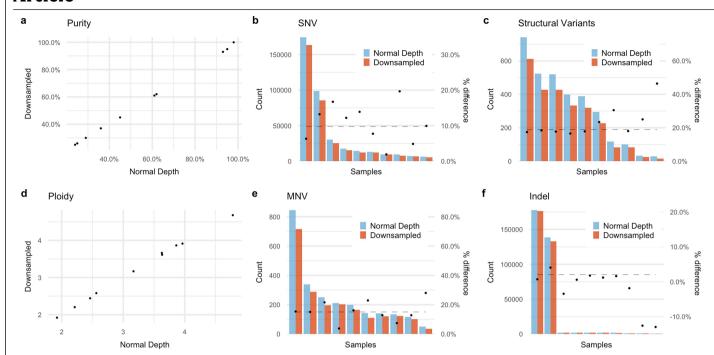
**Extended Data Fig. 3** | **SNV mutational signatures. a**, Prevalence and median mutational load of fitted COSMIC SNV mutational signature per cancer type (the number of patients per category is provided). The observed distribution largely reflects the patterns observed from primary cancers $^{13}$ . **b**, Box plots of relative residuals in fits per cancer type (sum of absolute difference between the fitted and actual divided by total mutational load). Boxes represent the twenty-fifth to seventy-fifth percentiles, and whiskers extend to the highest

and lowest values within 1.5× the upper/lower quartile distance, with outliers shown as dots.  $\mathbf{c}$ , Proportion of variants by 96 trinucleotide mutational context for two selected samples with high residuals and high mutational load. Top and bottom panels represent the highest outliers for breast (HMF002896) and oesophagus (HMF001562) cancers, respectively, from  $\mathbf{b}$ . Both of these samples were previously treated with the experimental drug SYD985—a duocarmycin-based HER2-targeting antibody—drug conjugate  $^{68}$ .



Extended Data Fig. 4 | Mutational load, genome-wide analyses and drivers. a, Proportion of samples by cancer type classified as microsatellite instable (MSIseq score > 4). b, Proportion of samples with a high mutational burden (TMB > 10 SNVs per Mb). c-e, Scatter plots of mutational load per sample for indels

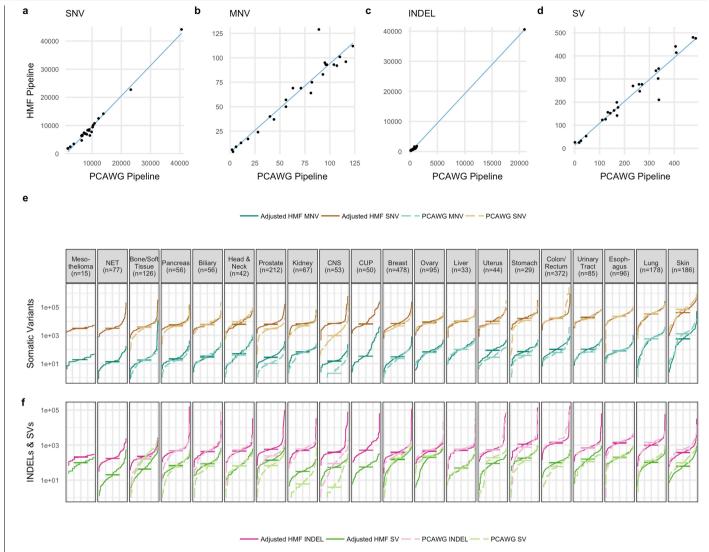
versus SNVs  $(\mathbf{c})$ , indels versus SVs  $(\mathbf{d})$ , and SVs versus SNVs  $(\mathbf{e})$ . MSI (MSIseq score > 4) and high TMB (>10 SNVs per Mb) thresholds are indicated.  $\mathbf{f}$ - $\mathbf{h}$ , Mean mutational load versus driver rate for SNVs  $(\mathbf{f})$ , indels  $(\mathbf{g})$  and SVs  $(\mathbf{h})$ , grouped by cancer type. MSI samples were excluded.



#### $Extended\,Data\,Fig.\,5\,|\,Effect\,of\,s equencing\,depth\,on\,variant\,calling.$

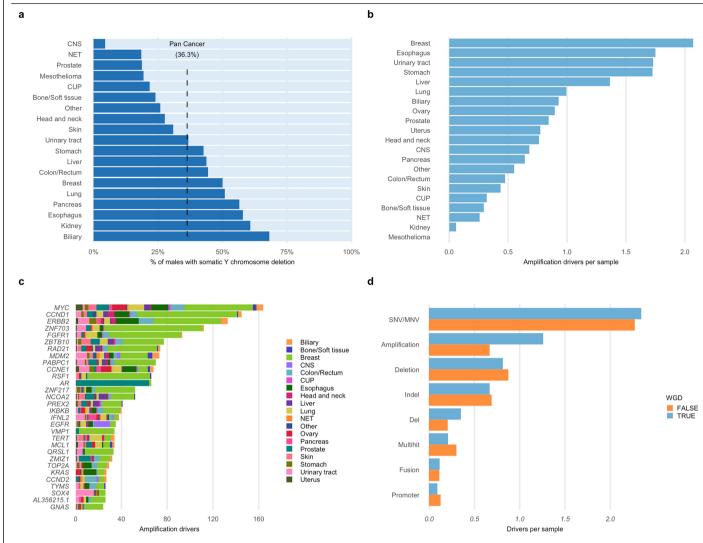
 $\mathbf{a}$ - $\mathbf{f}$ , Comparison of variant calling of ten randomly selected samples at normal depth and 50% downsampled (approximately 50 times, similar to the mean coverage for the PCAWG project<sup>14</sup>) for purity ( $\mathbf{a}$ ), SNV counts ( $\mathbf{b}$ ), SV counts ( $\mathbf{c}$ ),

ploidy ( $\mathbf{d}$ ), MNV counts ( $\mathbf{e}$ ) and indel counts ( $\mathbf{f}$ ). Decreasing coverage results in an average decrease in sensitivity of 10% for SNVs, 2% for indels, 15% for MNVs and 19% for SVs.



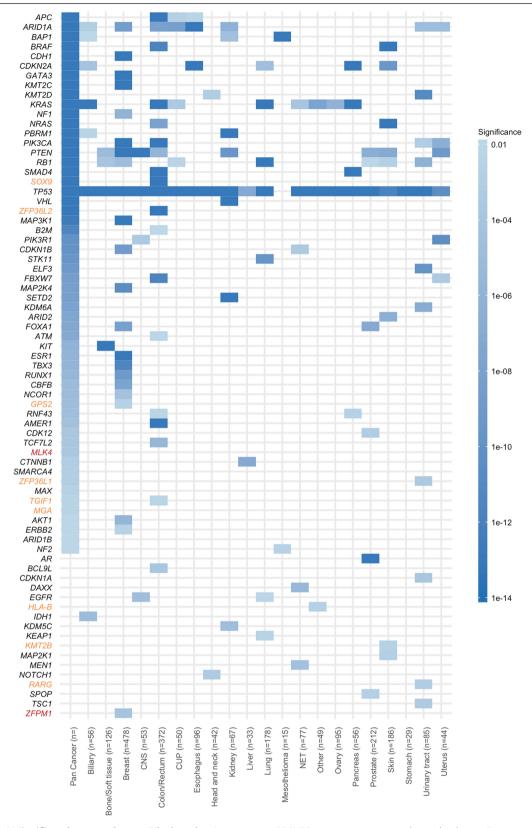
Extended Data Fig. 6 | Effect of bioinformatic analysis pipeline on variant calling. a-d, Comparison of observed mutational count per sample for SNVs (a), MNVs (b), indels (c) and SVs (d) on 24 patient samples analysed by the PCAWG and HMF pipelines. The PCAWG pipeline was found to have a 43% lower sensitivity for indels (which is based on a consensus calling), 18% lower for SVs (based on a different algorithm) and 6% lower for MNVs (only includes MNVs involving two nucleotides), with nearly the same sensitivity for SNVs. e, f, Cumulative distribution function plot for each tumour type (the number of independent patients per category is provided) of coverage and pipeline-adjusted mutational load for SNVs and MNVs (e) and indels and SVs (f). Mutational loads as shown in Fig. 1 were adjusted for the sensitivity effects

caused by differences in sequencing depth coverage (Extended Data Fig. 4) and analysis pipeline differences ( $\mathbf{a}$ – $\mathbf{d}$ ). After this correction, the TMB between primary and metastatic cohorts across all variant types are much more comparable ( $\mathbf{e}$ ,  $\mathbf{f}$ ), which indicates that technical differences do contribute to the reported mutational load differences between primary and metastatic tumours. Prostate cancer is the most notable exception, with approximately twice the TMB in all variant classes, although more subtle differences, potentially driven by biology, can also be observed for other tumour and mutation types. For cancer types that are comparable with the PCAWG cohort, the equivalent PCAWG numbers are shown by dotted lines. The median for each cohort is shown by a horizontal line.



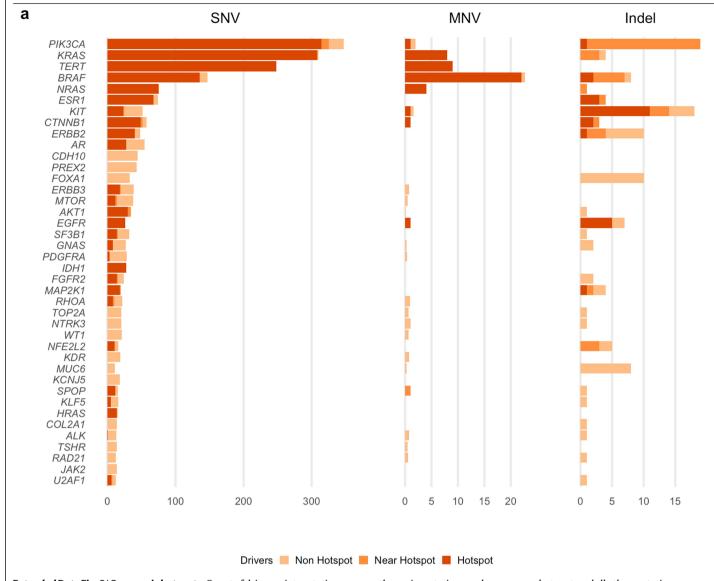
**Extended Data Fig. 7 | Somatic Y chromosome loss and driver amplifications. a**, Proportion of male tumours with somatic loss of more than 50% of Y chromosome (dark blue) grouped by tumour type. **b**, Mean rate of

amplification drivers per cancer type.  $\mathbf{c}$ , Breakdown of the number of amplification drivers per gene by cancer type.  $\mathbf{d}$ , Mean rate of drivers per variant type for samples with and without WGD.



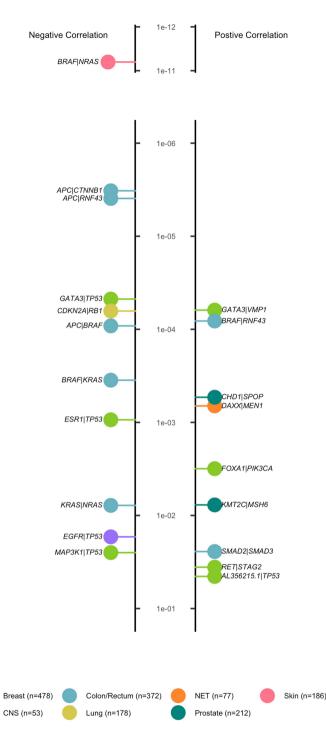
 $\label{lem:extended} \textbf{Data Fig. 8} | \textbf{Significantly mutated genes.} \ \text{Tile } \text{chart } \text{showing genes} \\ \text{found to be significantly mutated per cancer type (the number of independent patients per category is provided) and pan-cancer using dNdScv. Gene names marked in orange are also significant in a previous study$^{24}$, but not found in the $1.50 \times 10^{14}$.}$ 

 $COSMIC\ gene\ census\ or\ curated\ gene\ databases.\ Gene\ names\ marked\ in\ red\ are\ novel\ in\ this\ study.\ Significance\ (Poisson\ with\ Benjamini-Hochberg\ false\ discovery\ rate\ correction)\ is\ indicated\ by\ the\ intensity\ of\ shading.$ 



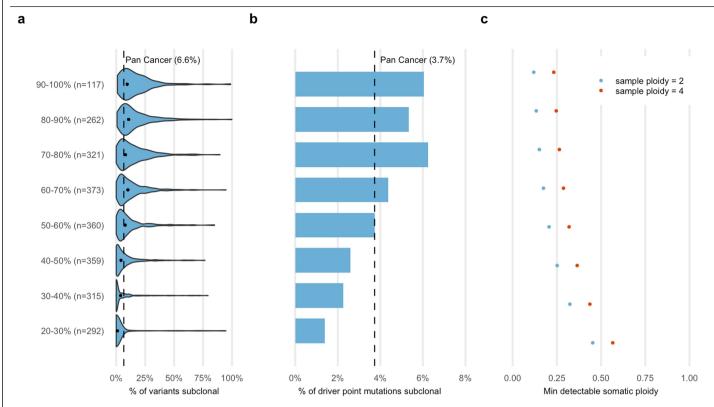
 $\textbf{Extended Data Fig. 9} | \textbf{Oncogenic hotspots.} \ Count of driver point mutations by variant type. Known pathogenic mutations curated from external databases are categorized as hotspot mutations. Mutations within five bases of a known mutations. The state of the$ 

pathogenic mutation are shown as near hot spot, and all other mutations are shown as non-hot spot.



**Extended Data Fig. 10** | **Driver co-occurrence. a**, Mutated driver gene pairs that are significantly positively (right) or negatively (left) correlated in individual tumour types (number of independent samples per tumour type is indicated in Fig. 1) sorted by q value (Fisher exact test adjusted for false discovery rate). Pairs of genes on the same chromosome that are frequently coamplified or co-deleted by chance are excluded from positively correlated results. The 20 significant findings include previously reported co-occurrence of mutated DAX-MEN1 in pancreatic NET  $(q=7\times10^{-4})$ , and CDH1-SPOP in prostate tumours  $(q=5\times10^{-4})$ , as well as negative associations of mutated genes within the same signal transduction pathway such as KRAS-BRAF  $(q=4\times10^{-4})$ 

and KRAS-NRAS (q=0.008) in colorectal cancer, BRAF-NRAS in skin cancer ( $q=6\times 10^{-12}$ ), CDKN2A-RBI in lung cancer ( $q=8\times 10^{-5}$ ) and APC-CTNNBI in colorectal cancer ( $q=3\times 10^{-6}$ ). APC is also strongly negatively correlated with both BRAF ( $q=9\times 10^{-5}$ ) and RNF43 ( $q=4\times 10^{-6}$ ), which together are characteristic of the serrated molecular subtype of colorectal cancers  $^{69}$ . SMAD2-SMAD3 are highly positively correlated in colorectal cancer (q=0.02), which supports a previous report in a large cohort of colorectal cancers  $^{70}$ . In breast cancer, we found several novel relationships, including a positive relationship for GATA3-VMPI ( $q=6\times 10^{-5}$ ) and FOXA1-PIK3CA ( $q=3\times 10^{-3}$ ), and a negative relationship for ESR1-TPS3 ( $q=9\times 10^{-4}$ ) and GATA3-TPS3 ( $q=5\times 10^{-5}$ ).



**Extended Data Fig. 11 | Subclonality of somatic variants. a**, Violin plot showing the percentage of point mutations per tumour purity bucket (the number of independent samples per category is indicated) that are subclonal in each purity bucket per sample. Black dots indicate the mean for each bucket. **b**, Percentage of driver point mutations that are subclonal in each purity bucket.

c, Approximate somatic ploidy detection cut-off of the HMF pipeline at median  $106 \times$  depth coverage for each purity bucket and for sample ploidy 2 and 4. Subclonal variants with cellular fraction less than this cut-off are unlikely to be detected by our pipeline analyses.



Corresponding author(s): Priestley and Cuppe
--

# **Reporting Summary**

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

## Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed
	The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
	The statistical test(s) used AND whether they are one- or two-sided  Only common tests should be described solely by name; describe more complex techniques in the Methods section.
	A description of all covariates tested
	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
	A full description of the statistics including <u>central tendency</u> (e.g. means) or other basic estimates (e.g. regression coefficient) AND <u>variation</u> (e.g. standard deviation) or associated <u>estimates of uncertainty</u> (e.g. confidence intervals)
$\boxtimes$	For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i> ) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted Give <i>P</i> values as exact values whenever suitable.
$\boxtimes$	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
$\boxtimes$	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
	$\boxtimes$ Estimates of effect sizes (e.g. Cohen's $d$ , Pearson's $r$ ), indicating how they were calculated
	Clearly defined error bars  State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on statistics for biologists may be useful.

## Software and code

Policy information about availability of computer code

Data collection

No software was used for data collection

Data analysis

All analyses are based on open source software, which is available from third parties or developed by Hartwig Medical Foundation and available on GitHub (https://github.com/hartwigmedical/). The table below lists all external and internally developed software/tools, versions used and public links to the source code.

External software/tools:

bcl2fastq 2.17 to 2.20 http://sapac.support.illumina.com/downloads/bcl2fastq-conversion-software-v2-20.html

BWA-mem 0.7.5a https://github.com/lh3/bwa

Sambamba 0.6.5 https://github.com/biod/sambamba/releases/tag/v0.6.5

Picard 1.141 https://broadinstitute.github.io/picard/

 ${\sf GATK~3.4.46~https://software.broadinstitute.org/gatk/download/auth?package=GATK-archive\&version=3.4-46-gbc02625}$ 

Strelka 1.0.14 https://github.com/Illumina/strelka

mutationalPatterns 1.4.3 https://bioc.ism.ac.jp/packages/3.6/bioc/html/MutationalPatterns.html

Manta 1.0.3 https://github.com/Illumina/manta

STAR-fusion ?? https://github.com/STAR-Fusion/STAR-Fusion/releases

Bioconductor CopyNumber package 1.24.0 http://bioconductor.org/packages/release/bioc/html/copynumber.html

ASCAT 2.52 https://github.com/Crick-CancerGenomics/ascat

dNdScv 0.1.0 https://github.com/im3sanger/dndscv/releases/tag/0.1.0

Circos 0.69.6 http://circos.ca/distribution/circos-0.69-6.tgz samtools 1.2 https://github.com/samtools/samtools/releases/tag/1.2 snpeff 4.3s https://sourceforge.net/projects/snpeff/files/snpEff\_v4\_3s\_core.zip/download vcftools 0.1.14 https://vcftools.github.io/index.html bcftools 1.9 https://github.com/samtools/bcftools/releases/download/1.9/bcftools-1.9.tar.bz2 HMF internal software/tools: Strelka\_post\_process 1.4 https://github.com/hartwigmedical/hmftools/releases/tag/strelka-post-process-v1-4 HMF pipeline v3.0 https://github.com/hartwigmedical/pipeline/releases/tag/v3.0 SAGE 1.1 https://github.com/hartwigmedical/hmftools/releases/tag/sage%E2%80%94v1-1 BPI 1.5 https://github.com/hartwigmedical/hmftools/releases/tag/bpi-v1-5 PURPLE 2.10 https://github.com/hartwigmedical/hmftools/releases/tag/purple-v2-10 Amber 1.5 https://github.com/hartwigmedical/hmftools/releases/tag/amber-v1-5 Cobalt 1.4 https://github.com/hartwigmedical/hmftools/releases/tag/cobalt-v1-4 healthchecker 2.1 https://github.com/hartwigmedical/hmftools/tree/master/health-checker R analysis suite 1.3 https://github.com/hartwigmedical/scripts/releases/tag/pancancerpaper-v1-3

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets

collection of clinical data from medical records.

- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data described in this study is freely available for academic use from the Hartwig Medical Foundation through standardized procedures and request forms which can be found at https://www.hartwigmedicalfoundation.nl/en/appyling-for-data/.

Available data includes germline and tumor raw sequencing data (BAM files, including non-mapped reads), annotated somatic and germline variants (VCF files with annotated SNV and indels, and pipeline output files for purity and ploidy status as well as copy number alteration and structural variants) and clinical data. Examples of output files can be found at https://resources.hartwigmedicalfoundation.nl. Briefly, a data request can be initiated by filling out the standard form in which intended use of the requested data is motivated. First, an advice on scientific feasibility and validity is obtained from experts in the field which is used as input by an independent Data Access Board who also evaluates if the intended use of the data is compatible with the consent given by the patients and if there would be any applicable legal or ethical constraints. Upon formal approval by the Data Access Board, a standard license agreement which does not have any restrictions regarding Intellectual Property resulting from the data analysis needs to be signed by an official organisation representative before access to the data is granted. After approval, access to data is provided under a license model, with the only main restriction that the data can only be used for the research detailed in the original request. Raw data files will be made available through a dedicated download portal with two-factor authentication.

Non-privacy sensitive somatic variants can also be browsed and explored through an open access web-based interface which can be accessed at http:// database.hartwigmedicalfoundation.nl/.

## Field-specific reporting

Please select the b	est fit for your research. If you are not sure, read the appropriate sections before making your selection.
\times Life sciences	Behavioural & social sciences Ecological, evolutionary & environmental sciences
For a reference copy of t	the document with all sections, see <u>nature.com/authors/policies/ReportingSummary-flat.pdf</u>
Life scier	nces study design
All studies must dis	sclose on these points even when the disclosure is negative.
Sample size	The metastatic tumor sample cohort described in the paper consists of 2520 independent samples from 2399 patients (including 121 repeat biopsies) collected in 41 hospitals (academic, teaching and general hospitals). No sample size calculations were performed as the main aim of the study was to build up a resource
Data exclusions	Samples that failed predefined QC criteria or with a tumor purity below 20% were excluded from all analyses and not included in the 2520 sample cohort (see Extended Data Fig 1). The tumor purity threshold was defined after bioinformatic tool optimization and simulations with titration series of reference samples and validation experiments on selected cohort samples.
Replication	Independent repeat processing of raw data of the same sample results in the same variant call data
Randomization	Not applicable as the primary goal of this study was to create a resource. The study and applyses do not include any experimental

manipulation and only involved the collection of tissue and blood material, the generation of whole genome sequencing data and the

Not applicable as the primary goal of this study was to create a resource. The study and analyses do not include any experimental manipulation and only involved the collection of tissue and blood material, the generation of whole genome sequencing data and the collection of clinical data from medical records.

# Reporting for specific materials, systems and methods

Materials & experimental systems		Methods	
n/a In	volved in the study	n/a	Involved in the study
	Unique biological materials	$\times$	ChIP-seq
$\boxtimes \Box$	Antibodies	$\boxtimes$	Flow cytometry
$\boxtimes   \square$	Eukaryotic cell lines	$\times$	MRI-based neuroimaging
$\boxtimes \Box$	Palaeontology		
$\boxtimes \Box$	Animals and other organisms		
	Human research participants		

## Unique biological materials

Policy information about availability of materials

Obtaining unique materials

Tumor biopsies are collected as part of two clinical studies and remaining material is deposited in local biobanks as described in the methods. Because of the nature of the material (very small amount), broad accessibility is not possible.

## Human research participants

Policy information about studies involving human research participants

Population characteristics

All patient included where diagnosed with metastatic disease and considered fit enough to undergo an invasive core-needle biopsy and planned to start treatment. The median age is 63 years (range 18 - 89). The cohort includes 1221 female and 1178 male subjects. Age and gender information of each patient is included in Supplementary Table 2. All patients were seen in hospitals in the Netherlands, including academic, teaching and general hospitals

Recruitment

Metastatic cancer patients were asked to participate in the studies in any of the 41 participating hospitals. Recruitment involved hundreds of medical specialists and research nurses which minimizes self-selection biases. Recruitment was independent on tumor type. An important requirement for participation was the ability to safely undergo a tumor biopsy. Health conditions and lesion site related risk could therefore have resulted in exclusion of patients.

# The clinical KRAS(G12C) inhibitor AMG 510 drives anti-tumour immunity

https://doi.org/10.1038/s41586-019-1694-1

Received: 29 March 2019

Accepted: 18 September 2019

Published online: 30 October 2019

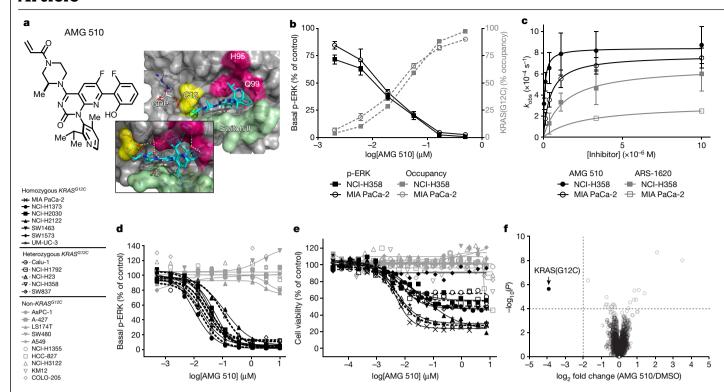
Jude Canon<sup>1\*</sup>, Karen Rex<sup>1,17</sup>, Anne Y. Saiki<sup>1,17</sup>, Christopher Mohr<sup>1</sup>, Keegan Cooke<sup>1</sup>, Dhanashri Bagal<sup>2</sup>, Kevin Gaida<sup>1</sup>, Tyler Holt<sup>1</sup>, Charles G. Knutson<sup>3</sup>, Neelima Koppada<sup>3</sup>, Brian A. Lanman<sup>1</sup>, Jonathan Werner<sup>1</sup>, Aaron S. Rapaport<sup>2</sup>, Tisha San Miguel<sup>1</sup>, Roberto Ortiz<sup>3,14</sup>, Tao Osgood<sup>1</sup>, Ji-Rong Sun<sup>1</sup>, Xiaochun Zhu<sup>3,15</sup>, John D. McCarter<sup>1</sup>, Laurie P. Volak<sup>3,16</sup>, Brett E. Houk<sup>4</sup>. Marwan G. Fakih<sup>5</sup>. Bert H. O'Neil<sup>6</sup>. Timothy J. Price<sup>7,8</sup>. Gerald S. Falchook<sup>9</sup>. Jayesh Desai<sup>10</sup>, James Kuo<sup>11</sup>, Ramaswamy Govindan<sup>12</sup>, David S. Hong<sup>13</sup>, Wenjun Ouyang<sup>2</sup>, Haby Henary<sup>4</sup>, Tara Arvedson<sup>2</sup>, Victor J. Cee<sup>1</sup> & J. Russell Lipford<sup>1\*</sup>

KRAS is the most frequently mutated oncogene in cancer and encodes a key signalling protein in tumours<sup>1,2</sup>. The KRAS(G12C) mutant has a cysteine residue that has been exploited to design covalent inhibitors that have promising preclinical activity<sup>3-5</sup>. Here we optimized a series of inhibitors, using novel binding interactions to markedly enhance their potency and selectivity. Our efforts have led to the discovery of AMG 510, which is, to our knowledge, the first KRAS(G12C) inhibitor in clinical development. In preclinical analyses, treatment with AMG 510 led to the regression of KRAS<sup>G12C</sup> tumours and improved the anti-tumour efficacy of chemotherapy and targeted agents. In immune-competent mice, treatment with AMG 510 resulted in a pro-inflammatory tumour microenvironment and produced durable cures alone as well as in combination with immune-checkpoint inhibitors. Cured mice rejected the growth of isogenic  $\mathit{KRAS}^{\mathit{G12D}}$  tumours, which suggests adaptive immunity against shared antigens. Furthermore, in clinical trials, AMG 510 demonstrated anti-tumour activity in the first dosing cohorts and represents a potentially transformative therapy for patients for whom effective treatments are lacking.

The KRAS oncoprotein is a GTPase and an essential mediator of intracellular signalling pathways that are involved in tumour cell growth and survival<sup>1,2</sup>. In normal cells, KRAS functions as a molecular switch, alternating between inactive GDP-bound and active GTP-bound states<sup>6,7</sup>. Transition between these states is facilitated by guanine nucleotide-exchange factors—which load GTP and activate KRAS—and GTP hydrolysis, which is catalysed by GTPase-activating proteins to inactivate KRAS<sup>2</sup>. GTP binding to KRAS promotes binding of effectors to trigger signal transduction pathways including the RAF-MEK-ERK (MAPK) pathway<sup>8,9</sup>. Somatic, activating mutations in KRAS are a hallmark of cancer and prevent the association of GTPase-activating proteins, thus stabilizing effector binding and enhancing KRAS signalling<sup>10</sup>. Although there are clinically approved inhibitors of several MAPK pathway proteins (for example, inhibitors of MEK, BRAF and EGFR) for a subset of tumour types, to date there have been no clinical molecules that are selective for KRAS-mutant tumours. Moreover, several MAPK-pathway-targeting therapies are contra-indicated for treatment of KRAS-mutant tumours owing to a lack of clinical efficacy<sup>11</sup>. Additionally, non-tumour or non-mutant selective  $the rapies \, can introduce \, on\text{-}target to xicities \, due to \, the \, inhibition \, of \, MAPK$ signalling in normal cells<sup>12,13</sup>. This might limit the ability to combine such agents with standard-of-care treatments or immunotherapy. Thus, there is a considerable unmet need for the development of tumour-selective therapies that do not introduce liabilities for normal cells.

KRAS<sup>G12C</sup> is present in approximately 13% of lung adenocarcinoma. 3% of colorectal cancer and 2% of other solid tumours 14. The mutant cysteine of KRAS(G12C) resides adjacent to a pocket (P2) that is present in the inactive GDP-bound form of KRAS<sup>3</sup>. The proximity of P2 and the mutant cysteine led to a broad search for covalent inhibitors, eventually resulting in the identification of ARS-1620<sup>3-5</sup>. This preclinical tool compound was a milestone for proof-of-concept, mutant-selective KRAS inhibition<sup>15</sup>. We identified a series of novel acrylamide-based molecules that utilize a previously unexploited surface groove in KRAS(G12C) to substantially enhance potency and selectivity. Intensive electrophile screening and structure-based design culminated in the discovery of AMG 510, which is, to our knowledge, the first KRAS(G12C) inhibitor to reach clinical testing in humans (clinical trials.gov identifier NCT03600883)<sup>16</sup>. Here we present the data on the preclinical activity of AMG 510, its ability to induce tumour-cell killing as monotherapy or when combined with other therapies, and the marked impact of AMG 510 on immune cell infiltration, which renders the tumour microenvironment highly sensitive to immunotherapy. We also present promising evidence for clinical efficacy.

1Amgen Research, Amgen Inc, Thousand Oaks, CA, USA. 2Amgen Research, Amgen Inc, South San Francisco, CA, USA. 3Amgen Research, Amgen Inc, Cambridge, MA, USA. 4Amgen Clinical Development, Amgen Inc, Thousand Oaks, CA, USA. 5City of Hope, Duarte, CA, USA. 6 Indiana University School of Medicine, Indianapolis, IN, USA. 7 The Queen Elizabeth Hospital, Woodville, South Australia, Australia. <sup>8</sup>University of Adelaide, Adelaide, South Australia, Australia. <sup>9</sup>Sarah Cannon Research Institute, Denver, CO, USA. <sup>10</sup>Peter MacCallum Cancer Center, Melbourne, Victoria, Australia. "Scientia Clinical Research, Randwick, New South Wales, Australia. 12 Washington University School of Medicine, St Louis, MO, USA, 13 The University of Texas MD Anderson Cancer Center, Houston, TX, USA. 14 Present address: Pfizer, La Jolla, CA, USA. 15 Present address: Takeda, Cambridge, MA, USA. 16 Present address: Celgene, San Diego, CA, USA. 17 These authors contributed equally: Karen Rex. Anne Y. Saiki. \*e-mail: icanon@amgen.com; ilipford@amgen.com



**Fig. 1**| **AMG 510** exploits a cryptic groove in KRAS(G12C) to enhance potency and selectivity. **a**, X-ray co-crystal structure of KRAS(G12C/C51S/C80L/C118S) bound to GDP and AMG 510 at a resolution of 1.65 Å. Cyan dashes, van der Waals contacts; white dashes, water-mediated interactions; yellow dashes, ligand-protein hydrogen bond interactions (PDB: 6OIM). **b**, Inhibition of p-ERK and occupancy of KRAS(G12C) by AMG 510 after a 2-h treatment. Data are mean  $\pm$  s.d., n = 3 replicates. **c**, Kinetic properties as determined by inhibition of p-ERK.  $k_{\rm obs}$  and standard error of the curve were determined from nonlinear

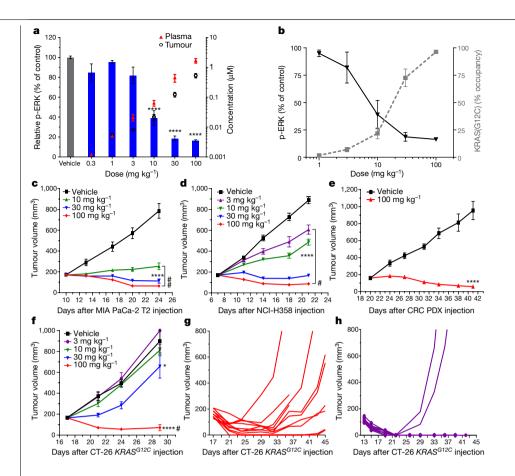
curve fitting of experimental values.  ${\bf d}$ ,  ${\bf e}$ , Cellular activity of AMG 510 across a panel of  $KRAS^{G12C}$  and non- $KRAS^{G12C}$  mutant cell lines measured as the inhibition of p-ERK after a 2-h treatment ( ${\bf d}$ ) and the effects on cell viability after a 72-h treatment ( ${\bf e}$ ). Representative examples of the data are shown (see Supplementary Table 1 for the number of replicates).  ${\bf f}$ , Cysteine proteome analysis of NCI-H358 whole-cell lysates after a 4-h treatment with 1  $\mu$ M AMG 510 or DMSO. n=5 independent replicates, P values were derived from a two-tailed Student's t-test.

#### Enhanced binding and potency of AMG 510

Direct inhibition of KRAS(G12C) was validated by ARS-1620, but the identification of improved inhibitors suitable for clinical testing has proven difficult. One key challenge is suboptimal potency owing to the small volume of the pocket occupied by ARS-1620, which offers limited avenues for additional protein-ligand interactions. This was illustrated by the X-ray crystal structure of the KRAS(G12C)-ARS-1620 covalent complex (Extended Data Fig. 1a), in which hydrogen bonding between ARS-1620 and His 95 featured prominently. Our key breakthrough was the discovery that a surface groove, created by an alternative orientation of His95, could be occupied by aromatic rings, which enhanced interactions with the KRAS(G12C) protein<sup>17</sup>. AMG 510 emerged as the top candidate from an optimization campaign of His95 groove-binding molecules, as it represented the convergence of improved potency and favourable development properties. The X-ray co-crystal structure of the covalent AMG 510-KRAS(G12C) complex (Fig. 1a and Extended Data Table 1) highlighted the binding of AMG 510 in the P2 pocket of KRAS. Although portions of the AMG 510 and ARS-1620 ligands are structurally related and overlap (Extended Data Fig. 1b), the His 95 groove is a novel feature of the binding of AMG 510 (Fig. 1a and Extended Data Fig. 1b). The highly optimized isopropyl-methylpyridine substituent of AMG 510 that occupied the His95 groove engaged in a continuous network of 25 ligand-protein van der Waals contacts extending from the backbone of helix 2 (His95, Tyr96) to the backbone of the flexible switch II loop (Fig. 1a). These enhanced interactions improved the potency of AMG 510 approximately 10-fold (mean half-maximum inhibitory concentration (IC<sub>50</sub>) =  $0.09 \,\mu\text{M}$ ), as compared to ARS-1620 in a nucleotide-exchange assay with recombinant GDP-bound KRAS(G12C). AMG 510 did not inhibit wild-type KRAS and a non-reactive analogue did not inhibit KRAS(G12C) (Extended Data Fig. 1c, d). The kinetics of the reaction between AMG 510 and GDP-KRAS(G12C) were measured by mass spectrometry and exhibited a marked improvement compared to ARS-1620 (Extended Data Fig. 1e, f). Relative to cysteine-targeted kinase inhibitors in the clinic<sup>18</sup>, AMG 510 exhibited a larger maximal rate of inactivation ( $k_{\text{inact}}$ ), consistent with the KRAS-induced catalysis mechanism that has previously been described for ARS-1620<sup>15</sup>. The non-specific reactivity of AMG 510 with glutathione was relatively slow ( $t_{1/2}$  = 196 min)<sup>19</sup> and within the range of clinical acrylamides<sup>20</sup>.

## AMG 510 inhibits signalling and growth

The cellular activity of AMG 510 was assessed by measuring basal phosphorylation of ERK1/2 (p-ERK) and by mass spectrometry to detect the covalent conjugation or occupancy of KRAS(G12C) by AMG 510. In two KRAS<sup>G12C</sup> cell lines, NCI-H358 and MIA PaCa-2, AMG 510 almost completely inhibited p-ERK ( $IC_{50} \approx 0.03 \,\mu\text{M}$ ) after a 2-h treatment and was 20-fold more potent than ARS-1620 (Extended Data Fig. 1g). This inhibition closely tracked the occupancy of KRAS(G12C) by AMG 510, with near maximal levels achieved in both assays at around  $0.2\,\mu\text{M}$  (Fig. 1b). AMG 510 also potently impaired cellular viability in both NCI-H358 and MIA PaCa-2 (IC $_{50} \approx 0.006 \, \mu M$  and 0.009 µM respectively, approximately 40-fold more potent than ARS-1620;  $Extended \, Data \, Fig. \, 1h). \, Examining \, the \, time \, and \, concentration \, dependence \, and \, concentration \, depen$ of the inhibition of p-ERK in these lines revealed a kinetic advantage that favoured AMG 510 by approximately 22-fold (Fig. 1c and Extended Data Fig. 1f). The maximal inhibition rate of p-ERK by AMG 510 is approximately twofold greater than the rate-limiting GTP-KRAS(G12C) hydrolysis rate that has recently been proposed<sup>4</sup>. To estimate the GTPase rate by another method, we used a SHP2 inhibitor<sup>21</sup> to eliminate all upstream signalling to KRAS, which yielded a rate  $(9.4 \times 10^{-4} \,\mathrm{s}^{-1}, t_{1/2} = 12.2 \,\mathrm{min}; \,\mathrm{Extended \, Data})$ Fig. 2a) that was congruent with what was observed for AMG 510.



and growth of KRAS<sup>G12C</sup>-mutant tumours in vivo. a, p-ERK levels in NCI-H358 tumours 2 h after a single dose of vehicle (black bar) or AMG 510 (blue bars). AMG 510 concentrations in plasma (red triangles) or tumours (black open circles). Data are mean  $\pm$  s.e.m., n = 3 mice per group; \*\*\*\*P < 0.0001; one-way analysis of variance (ANOVA) followed by Dunnett's multiple-comparison test. b, AMG 510 treatment results in covalent modification of KRAS(G12C). which inversely correlates with p-FRK inhibition in NCI-H358 tumours. Data are mean  $\pm$  s.d., n = 3 mice per group.  $\mathbf{c} - \mathbf{f}$ , Mice with established MIA PaCa-2T2 tumours (c), NCI-H358 tumours (d), colorectal-cancer patient-derived xenografts (CRC PDX; e) or CT-26 KRAS<sup>G12C</sup> tumours (f) were treated with AMG 510. Data are mean  $\pm$  s.e.m., n = 10 mice per group, except for  $\mathbf{e}$ , n = 8 mice per group; \*\*\*\*P<0.0001,\*P<0.05 compared with vehicle; repeated-measures ANOVA followed by Dunnett's multiple-comparison test;  ${}^{\#}P < 0.05$ regression by two-sided Student's t-test. g-h, Individual CT-26 KRAS<sup>G12C</sup> tumour volume plots from mice treated with 100 mg kg<sup>-1</sup> AMG 510 (g) or 200 mg kg<sup>-1</sup> AMG 510 (**h**) (n = 10 per group).

Fig. 2 | AMG 510 inhibits ERK phosphorylation

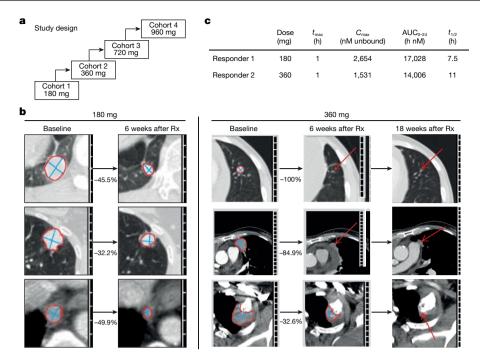
For further evaluation of the signalling effect of KRAS(G12C) inhibition, two cell lines were treated with a titration of AMG 510 for 4 or 24 h, and signalling nodes were analysed by immunoblot (Extended Data Fig. 2b). The KRAS species shifted mobility upon the formation of covalent adducts with AMG 510 and accumulated with increasing time and dose, consistent with downstream inhibition of the MAPK pathway (that is, p-MEK1/2 and p-ERK1/2) in both cell lines (Extended Data Fig. 2b). KRAS(G12C) inhibition by AMG 510 also led to an accumulation of active EGFR (p-EGFR(Y1068)). Inhibition of AKT phosphorylation (p-AKT) was apparent in one cell line, whereas a decrease in S6 phosphorylation (p-S6) and an increase in cleaved caspase-3 were observed at 24 h in both lines, suggesting induction of apoptosis. In time course studies. treatment with AMG 510 at 0.1 µM (Extended Data Fig. 2c) elicited rapid (<2 h) and sustained (>24 h) effects on MAPK and EGFR pathway signalling, whereas p-S6 and caspase cleavage emerged 8-16 h after treatment in both lines. To assess activity and selectivity, AMG 510 was profiled in 22 cell lines that had heterozygous or homozygous KRAS<sup>G12C</sup>, KRAS mutations other than KRAS<sup>G12C</sup> or wild-type KRAS. Treatment with AMG 510 for  $2\,h\,showed\,that\,basal\,p\text{-ERK}\,was\,inhibited\,in\,all\,\textit{KRAS}^{\textit{GI2C}}\,cell\,lines, with$  $IC_{50}$  values ranging from 0.010  $\mu$ M to 0.123  $\mu$ M (Fig. 1d and Supplementary Table 1). AMG 510 did not inhibit p-ERK in any of the non-KRAS<sup>G12C</sup> lines (IC<sub>50</sub> > 10  $\mu$ M; Fig. 1d and Supplementary Table 1). In cell-viability assays, AMG 510 impaired the growth of all KRASG12C cell lines, except SW1573, with IC<sub>50</sub> values ranging from  $0.004 \mu M$  to  $0.032 \mu M$  (Fig. 1e and Supplementary Table 1). Non-KRASG12C lines were insensitive to AMG 510 (IC<sub>50</sub> > 7.5  $\mu$ M; Fig. 1e and Supplementary Table 1). As reported for other KRAS(G12C) inhibitors<sup>4,5</sup>, spheroid growth conditions enhanced the sensitivity of most tested lines to AMG 510 (Extended Data Fig. 2d and Supplementary Table 1). To further determine the selectivity of the covalent interaction of AMG 510 with KRAS(G12C) and to identify other potential 'off-target' cellular proteins, cysteine-proteome profiling by mass spectrometry was performed as previously described<sup>4</sup>. After 4-h treatment with DMSO or 1 µM AMG 510 (>30-fold above p-ERK IC<sub>50</sub>), the

cysteine proteome was enriched and peptides were identified. Among 6,451 unique cysteine-containing peptides, the Cys12 peptide from KRAS(G12C) was the only peptide that met the criteria for covalent target engagement<sup>4</sup> (Fig. 1f and Supplementary Table 2).

The effect of AMG 510 treatment on KRAS(G12C) signalling in vivo was evaluated in pharmacodynamics assays in which p-ERK was measured. In three  $\mathit{KRAS}^{\mathit{G12C}}$  tumour models, AMG 510 inhibited p-ERK in a dose-dependent manner 2 h after treatment (Fig. 2a and Extended Data Fig. 3a, b) and maximal inhibition was observed at 30–100 mg kg<sup>-1</sup>. Time-course pharmacodynamics assays demonstrated peak plasma and tumour exposure of AMG 510 0.5 h after a single dose (10 mg kg<sup>-1</sup>), leading to maximal inhibition of p-ERK 2-4 hafter treatment and sustained inhibition for 48 h (Extended Data Fig. 3c, d). This was consistent with covalent inhibition of the long-lived KRAS(G12C) protein ( $t_{1/2} \approx 20-24$  h; Extended Data Fig. 3e). Occupancy of KRAS(G12C) by AMG 510 was also measured by mass spectrometry and approached 100% at 100 mg kg<sup>-1</sup>, correlating with maximal suppression of p-ERK (Fig. 2b and Extended Data Fig. 3f). Time-course studies indicated that occupancy was detected by 0.5 h and maximal at 2 h (Extended Data Fig. 3g).

#### Mutant-selective tumour inhibition in vivo

In mice with xenografts of human tumour cells, AMG 510 significantly inhibited the growth of MIA PaCa-2T2 and NCI-H358 tumours at all doses, and regression of tumours was observed at higher doses (Fig. 2c, d). The dose of AMG 510 that was required to achieve the regression of MIA PaCa-2T2 tumours was at least 3.3-fold lower than ARS-1620 (Extended Data Fig. 4a). Plasma exposures above the cellular IC<sub>90</sub> of p-ERK for more than 2 h resulted in tumour regression (Extended Data Fig. 4b, c). AMG 510 also inhibited the growth of *KRAS*<sup>GI2C</sup>-mutant patient-derived xenografts (Fig. 2e and Extended Data Fig. 4d). By contrast, AMG 510 treatment had no effect on KRAS<sup>G12V</sup> tumour growth (Extended Data Fig. 4e). In immune-competent mice, AMG 510 resulted in regression of CT-26



**Fig. 3** | **Clinical activity of AMG 510 in patients with lung cancer in first-in-human dose-escalation study. a**, Study design. **b**, Computed tomography scans of patients with  $KRAS^{GIZC}$  lung carcinoma treated with AMG 510 (left, 180 mg; right, 360 mg). Representative pre-treatment (baseline) and post-treatment ( $R_x$ ) scans. Lesions are outlined by a red outline or highlighted by red

arrows. Left images show the lower-right lobe of the lungs (top), upper-left lobe of the lungs (middle) and lymph node (bottom). Right images show the upper left lobe of the lungs (top) and pleura (middle and bottom). Lesions in the 18-week scans of the patient who received 360 mg AMG 510 were considered too small to accurately measure.  ${\bf c}$ , Pharmacokinetic data from the two responders.

 $\it KRAS^{\it G12C}$  tumours, a mouse syngeneic tumour model generated by CRISPR technology (Fig. 2f). Two of the ten mice in the 100 mg kg<sup>-1</sup> group had no detectable tumours at the end of the study (day 29). However, regression of the tumours lacked durability (Fig. 2g), possibly owing to incomplete inhibition of p-ERK (Extended Data Fig. 3b). Therefore, a dose of 200 mg kg<sup>-1</sup> of AMG 510 was evaluated, resulting in near-complete inhibition of p-ERK (Extended Data Fig. 3b) and durable cures in eight out of ten mice (Fig. 2h), in which AMG 510 plasma levels were just below the cellular IC $_{90}$  (Extended Data Fig. 4f). Intriguingly, in the same tumour model but in mice that lacked T cells, AMG 510 induced regression but not cures, suggesting that the immune system drives cures in immunecompetent mice (Extended Data Fig. 4g).

#### **Evidence of clinical activity**

The enhanced potency and efficacy of AMG 510 prompted its selection as, to our knowledge, the first KRAS(G12C) inhibitor to enter clinical trials (clinicaltrials.gov identifier NCT03600883)<sup>16</sup>. AMG 510 was administered orally, once daily, in escalating dosing cohorts (Fig. 3a). In the first two dosing cohorts there were four patients with non-small-cell lung carcinoma (180 mg, n=3; 360 mg, n=1). Treatment with AMG 510 resulted in objective partial responses (as per RECIST 1.1) in two patients (Fig. 3b and Extended Data Fig. 5) and stable disease in two patients. The two patients with a partial response had progressed on multiple previous systemic treatments including carboplatin, pemetrexed and nivolumab with documented disease progression. After 6 weeks of treatment with AMG 510, the first responder (180 mg) exhibited tumour shrinkage of 34%, and the second (360 mg) exhibited a tumour reduction of 67%. A follow-up scan at 18 weeks revealed complete resolution of target lesions in the second responder. AMG 510 exposures in both patients were above the cellular IC<sub>90</sub> of p-ERK (165 nM in MIA PaCa-2; Extended Data Fig. 4b) for 24 h (Fig. 3c). These patients remain active on AMG 510 treatment with the durations of 42 and 29 weeks, respectively, as of the cut-off date for the present data. We show that these patients responded to a mutant-specific KRAS inhibitor, representing a milestone for patients with KRAS<sup>GI2C</sup>-mutant cancer.

#### AMG 510 improves efficacy of targeted agents

The clinically validated strategy of combining BRAF and MEK inhibitors<sup>22</sup> suggests that the combinations of AMG 510 and other inhibitors in the MAPK (and AKT) signalling pathways might enhance tumourcell killing and overcome resistance. Therefore, in vitro combination experiments were conducted in several KRAS<sup>G12C</sup> cell lines with matrices of AMG 510 and inhibitors of HER kinases, EGFR, SHP2, PI3K, AKT and MEK (Extended Data Fig. 6a and Supplementary Table 3). As suggested by the induction of p-EGFR by AMG 510 (Extended Data Fig. 2b), the combination of AMG 510 with multiple agents resulted in synergistic killing<sup>23</sup> of NCI-H358 tumour cells (Fig. 4a, Extended Data Fig. 6a and Supplementary Table 3). Synergy was more limited in other lines, but the combination with a MEK inhibitor was synergistic in multiple settings and was enhanced in spheroid growth conditions (Fig. 4a). Significantly enhanced anti-tumour activity was also observed in vivo with a minimally efficacious dose of AMG 510 in combination with a MEK inhibitor, when compared to either of the single agents alone (Fig. 4b). These data suggest that the clinical combination of AMG 510 with MAPK inhibitors might eliminate bypass or residual signalling that could limit efficacy or induce resistance.

Given the prevalence of *KRAS*<sup>GI2C</sup> in lung adenocarcinoma, a combination treatment of AMG 510 with carboplatin, a standard-of-care chemotherapeutic, was investigated. Treatment with either AMG 510 or carboplatin resulted in significant inhibition of NCI-H358 tumour growth in mice (Fig. 4c). However, combination treatment at various doses resulted in significantly improved anti-tumour efficacy (Fig. 4c and Extended Data Fig. 6b). The demonstration of enhanced efficacy of the combination of a mutant-selective KRAS inhibitor and a chemotherapeutic agent provides rationale for this approach in the clinic.

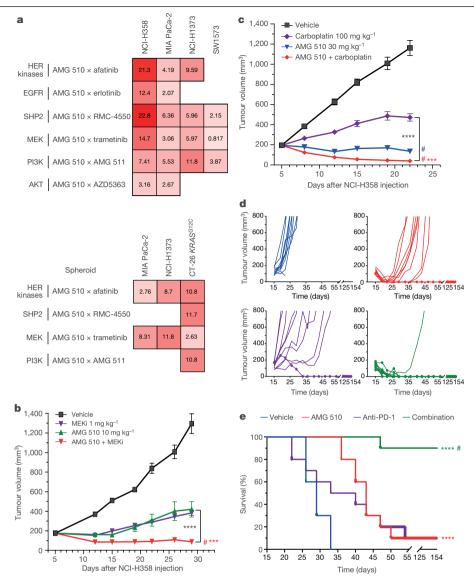


Fig. 4 | AMG 510 combined with cytotoxic or targeted agents results in enhanced efficacy. a, Synergy scores for AMG 510 combinations with targeted agents represented as a heat map, with higher scores (darker red) denoting stronger synergy. **b**, AMG 510 in combination with a MEK inhibitor (PD-0325901). c, AMG 510 in combination with carboplatin. d, CT-26 KRAS<sup>G12C</sup> tumour growth in individual mice. Lines with circles indicate tumour-free mice. e, Kaplan-Meier analysis of survival end point (tumour size > 800 mm<sup>3</sup>). b, c, Data are

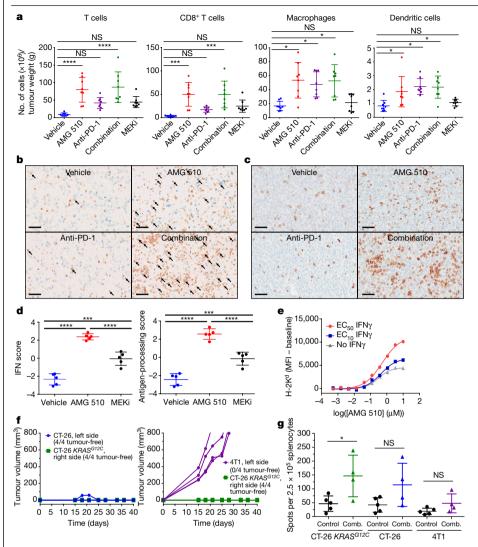
mean  $\pm$  s.e.m., n = 10 mice per group; \*\*\* $P \le 0.001$  for combination treatment compared with the single agent; \*\*\*\*P < 0.0001 for treatment group compared with vehicle; P values were determined by repeated-measures ANOVA followed by Dunnett's multiple-comparison test; #P < 0.001 regression by two-sided Student's *t*-test. **e**, n = 10 mice per group; \*\*\*\*P < 0.0001 compared with vehicle;  $^{\#}P < 0.001$  combination versus AMG 510 or anti-PD-1 determined by two-sided Mantel-Cox test.

## AMG 510 synergizes with immunotherapy

Blockade of the immune checkpoint axis that involves programmed cell death1(PD-1)-programmed death ligand1(PD-L1) is clinically validated in multiple settings. As the long term cures induced by AMG 510 in the CT-26 KRAS<sup>G12C</sup> model were dependent on the engagement of the immune system (Fig. 2h and Extended Data Fig. 4g), strategies such as anti-PD-1 therapy that further boost anti-tumour T cell activity may synergize with AMG 510. The CT-26  $KRAS^{G12C}$  model is dependent on the  $KRAS^{G12C}$  allele (Extended Data Fig. 7a, b) and is sensitive to AMG 510 treatment (Fig. 2f and Extended Data Fig. 3b). Furthermore, its parental line CT-26 has been broadly used to evaluate the effects of immunotherapy as well as combinations of immunotherapeutic and targeted agents<sup>24–26</sup>. Therefore, we used this model to evaluate the combination of anti-PD-1 immune checkpoint inhibition with AMG 510, which was administered at a suboptimal dose to enable the evaluation of combination effects. As shown above (Fig. 2f, g), AMG 510 caused tumour regression in mice as a single agent (Fig. 4d),

but only one out of ten tumours remained completely regressed (Fig. 4d). Anti-PD-1monotherapy delayed tumour growth, with complete regression in only one of ten tumours. Notably, combined treatment led to complete responses in nine out of ten mice (Fig. 4d). Treatment was stopped after day 43, and all complete responders remained cured 112 days later. Using a surrogate end point (tumour volume > 800 mm<sup>3</sup>), the combined treatment significantly improved survival (Fig. 4e).

To understand the effects of treatment on immune cell composition, CT-26 KRAS GIZC tumours were immunophenotyped. After 4 days of treatment, AMG 510 markedly increased the infiltration of T cells, primarily CD8+T cells, into the tumour (Fig. 5a and Extended Data Fig. 8a). Increased infiltration of CD8+T cells was also observed in the combination group, but not after anti-PD-1 monotherapy. Immunohistochemical analysis also revealed an increased number of total and proliferating CD3<sup>+</sup>T cells and total CD8<sup>+</sup>T cells after AMG 510 treatment, which were further increased after the combination treatment (Fig. 5b, c). As an



**Fig. 5** | **AMG 510 treatment induces a proinflammatory tumour microenvironment. a**-**c**, CT-26 *KRAS*<sup>CI2C</sup> tumours were immunophenotyped by flow cytometry (**a**) or immunohistochemistry of CD3 and Ki-67 (**b**) or CD8 (**c**).

 $\mathbf{a}$ , n = 8 per group;  $\mathbf{b}$ ,  $\mathbf{c}$ , n = 5 per group. Ki-67 is shown in blue and has a nuclear localization: CD3 and CD8 stains are shown in brown and are cytoplasmic. Arrows in b highlight examples that are doublepositive for CD3 and Ki-67. Scale bars, 50 µm. d, RNA was isolated from CT-26 KRAS<sup>G12C</sup> tumours (n = 5 per group). Gene expression and scores were calculated by NanoString technology, e. Cell surface expression of MHC class I antigen H-2Kd on CT-26 KRASG12C cells that were treated with AMG 510 together with or without IFN y as measured by flow cytometry. f, Individual tumour growth in mice that were cured with AMG 510 and anti-PD-1 treatment and were then rechallenged with CT-26 KRAS GIZC, CT-26 or 4T1 cells. g, Splenocytes were collected and challenged with CT-26, CT-26 KRAS<sup>G12C</sup> or 4T1 cells. Levels of secreted IFNy were measured by ELISpot assay. n = 5 control, n=4 combination. **a**, **d**, **g**, Data are mean  $\pm$  s.d. **a**, **d**, \*\*\*\*P < 0.0001. \*\*\*P < 0.001. \*P < 0.05: NS. not significant; one-way ANOVA followed by Tukey's test.  $\mathbf{g}$ , \*P = 0.0269; two-sided Student's t-test.

additional comparison we used a MEK inhibitor, which blocked MAPK signalling downstream of RAS (Extended Data Fig. 8b). This inhibitor regressed CT-26  $KRAS^{G12C}$  tumours in mice to a similar level as AMG 510 (Extended Data Fig. 8c, d), but did not significantly affect the numbers of infiltrating CD8+T cells (Fig. 5a). AMG 510 treatment also increased the infiltration of macrophages and dendritic cells, including CD103+ cross-presenting dendritic cells, which are critical for T cell priming and activation and are implicated in T cell recruitment  $^{27}$  (Fig. 5a and Extended Data Fig. 8a). PD-1 expression on CD8+T cells was moderately increased by both AMG 510 and the MEK inhibitor (Extended Data Fig. 8a).

Tumour RNA was purified after 2 days of treatment for transcriptional profiling of a panel of immune-associated genes. AMG 510 induced a pro-inflammatory microenvironment characterized by increased interferon signalling, chemokine production, antigen processing, cytotoxic and natural killer cell activity, as well as markers of innate immune system stimulation, that were significantly higher compared to the effects induced by MEK inhibition (Fig. 5d and Extended Data Fig. 8e). Infiltration of immune cells was correlated with increased expression of several chemokines including Cxcl11 (Extended Data Fig. 8e and Supplementary Table 4). To examine whether these immune-enhancing effects were directly attributable to AMG 510, CT-26 KRAS<sup>G12C</sup> cells were treated with AMG 510 in vitro and the expression of immune genes was measured. AMG 510 induced expression of Cxcl10 and Cxcl11 (Extended Data Fig. 9a), which are key attractants of tumour-suppressive immune cells<sup>27,28</sup>. This provides a potential mechanistic link by which AMG 510 treatment increases the intratumoral concentration of chemokines, leading to the infiltration of T cells and dendritic cells and improved immunosurveillance.

Previous data suggested that although MEK inhibition could promote anti-tumour activity in combination with anti-PD-L1 treatment in vivo, it can also inhibit T cell function<sup>24</sup>. Using an in vitro co-culture system with mouse bone marrow-derived dendritic cells and transgenic CD8<sup>+</sup>T cells, MEK inhibition impaired antigen-specific T cell proliferation, whereas AMG 510 did not affect the T cell response (Extended Data Fig. 9b). Furthermore, AMG 510 induced expression of MHC class lantigens on CT-26 KRAS<sup>G12C</sup> tumour cells in vitro (Fig. 5e and Extended Data Fig. 9c). These data suggest that AMG 510 treatment leads to increased T cell priming, antigen recognition of tumour cells and the potential establishment of long-term anti-tumour T cell responses. To test this, mice that were cured by the combined treatment of AMG 510 and anti-PD-1 (Fig. 4d) were rechallenged with bilateral tumours of CT-26 KRAS<sup>G12C</sup> and parental CT-26 (KRAS<sup>G12D</sup>) cells, or CT-26 KRAS<sup>G12C</sup> and an unrelated mouse breast tumour model, 4T1. All 4T1 tumours (four out of four) grew, but none of the CT-26 KRAS<sup>G12C</sup> tumours (zero out of eight) or CT-26 parental tumours (zero out of four) became established (Fig. 5f). In a control group of naive mice, all parental CT-26 and CT-26 KRAS<sup>G12C</sup> tumours grew (15 out of 15; Extended Data Fig. 9d). Splenocytes collected from the cured mice were stimulated with CT-26, CT-26 KRAS<sup>G12C</sup> or 4T1 tumour cells, and we measured secreted IFNy as a marker of tumour-specific T cell priming and activity. CT-26 KRAS GIZC cells and parental CT-26 cells caused nearly a threefold increase in IFNγ, which was not induced by 4T1 cells (Fig. 5g). Together, these data suggest that the combination of AMG 510 and anti-PD-1 therapy prompted the establishment of long-term tumour-specific T cell responses.

#### Discussion

The discovery of the interaction with the His95 groove of KRAS(G12C) enabled markedly increased potency and the identification of AMG 510. a first-in-class oral KRAS(G12C) inhibitor with evidence of clinical activity in patients with KRAS<sup>G12C</sup> mutant cancer. Preclinically, AMG 510 selectively targeted KRAS<sup>G12C</sup> tumours, caused durable regression as a monotherapy, and could be combined with cytotoxic and targeted agents to synergistically kill tumour cells. AMG 510 treatment led to an inflamed tumour microenvironment that was highly responsive to immune-checkpoint inhibition. Combined treatment of anti-PD-1 therapy and a MEK inhibitor has shown preclinical efficacy in several reports<sup>24,29,30</sup>, and this was associated with increased T cell infiltration. In the present study, significantly greater immune cell infiltration was observed after selective KRAS(G12C) inhibition compared to the MEK inhibitor. In contrast to the reported effects of non-tumour-selective MEK inhibition, which blocks T cell expansion and priming<sup>24</sup>, selective inhibition of KRAS(G12C) by AMG 510 resulted in increased T cell infiltration and activation. Furthermore, the combination of AMG 510 and anti-PD-1 therapy established a memory T cell response against both the CT-26 KRAS<sup>G12C</sup> cells and the parental CT-26 tumour cells. These data support a model of enhanced antigen recognition and T cell memory in which AMG 510-induced tumour cell death and innate immune responses, combined with anti-PD-1 treatment, results in an adaptive immune response that can recognize and eradicate related but non- $KRAS^{G12C}$  tumours. There is ample evidence that the intratumoral KRASmutation status can be heterogeneous within the same tumour and between primary and metastatic sites  $^{31-33}$ . Taken together, our data suggest that AMG 510 might be an effective anti-tumour agent even in settings in which KRAS<sup>G12C</sup> expression is heterogenous.

#### **Reporting summary**

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

## **Data availability**

Most of the data generated or analysed during this study are included in this published Article or available as Source Data. X-ray crystallographic coordinates and structure factor files have been deposited in the Protein Data Bank (PDB: 60IM). Other data that support the findings of this study are available from the corresponding authors. Qualified researchers may request data from Amgen clinical studies. Further details are available at http://www.amgen.com/datasharing.

#### **Online content**

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1694-1.

- Barbacid, M. ras genes. Annu. Rev. Biochem. 56, 779-827 (1987).
- Simanshu, D. K., Nissley, D. V. & McCormick, F. RAS proteins and their regulators in human disease. Cell 170, 17-33 (2017)
- Ostrem, J. M., Peters, U., Sos, M. L., Wells, J. A. & Shokat, K. M. K-Ras(G12C) inhibitors allosterically control GTP affinity and effector interactions. Nature 503, 548-551 (2013).

- Patricelli, M. P. et al. Selective inhibition of oncogenic KRAS output with small molecules targeting the inactive state. Cancer Discov. 6, 316-329 (2016).
- Janes, M. R. et al. Targeting KRAS mutant cancers with a covalent G12C-specific inhibitor. Cell 172, 578-589 (2018).
- Pai, E. F. et al. Structure of the guanine-nucleotide-binding domain of the Ha-ras oncogene product p21 in the triphosphate conformation, Nature 341, 209-214 (1989).
- Milburn, M. V. et al. Molecular switch for signal transduction: structural differences between active and inactive forms of protooncogenic ras proteins. Science 247, 939-945
- Cully, M. & Downward, J. SnapShot: Ras signaling. Cell 133, 1292-1292.e1 (2008)
- Vetter, I. R. & Wittinghofer, A. The guanine nucleotide-binding switch in three dimensions. Science 294 1299-1304 (2001)
- Scheffzek, K. et al. The Ras-RasGAP complex: structural basis for GTPase activation and its loss in oncogenic Ras mutants. Science 277, 333-338 (1997).
- Jimeno, A., Messersmith, W. A., Hirsch, F. R., Franklin, W. A. & Eckhardt, S. G. KRAS mutations and susceptibility to cetuximab and panitumumab in colorectal cancer. Cancer J. 15, 110-113 (2009).
- Welsh, S. J. & Corrie, P. G. Management of BRAF and MEK inhibitor toxicities in patients with metastatic melanoma, Ther. Adv. Med. Oncol. 7, 122-136 (2015).
- Fakih, M. & Vincent, M. Adverse events associated with anti-EGFR therapies for the treatment of metastatic colorectal cancer, Curr, Oncol. 17, S18-S30 (2010)
- AACR Project GENIE Consortium. AACR Project GENIE: powering precision medicine through an international consortium. Cancer Discov. 7, 818-831 (2017).
- Hansen, R. et al. The reactivity-driven biochemical mechanism of covalent KRAS<sup>G120</sup> inhibitors, Nat. Struct, Mol. Biol. 25, 454-462 (2018).
- clinicaltrials.gov. A Phase 1/2, Study Evaluating the Safety, Tolerability, PK, and Efficacy of AMG 510 in Subjects With Solid Tumors With a Specific KRAS Mutation https://clinicaltrials. gov/ct2/show/NCT03600883 (2018).
- Gentile, D. R. et al. Ras binder induces a modified switch-II Pocket in GTP and GDP states. Cell Chem. Biol. 24, 1455-1466 (2017).
- Schwartz, P. A. et al. Covalent EGFR inhibitor analysis reveals importance of reversible interactions to potency and mechanisms of drug resistance. Proc. Natl Acad. Sci. USA 111,
- Cee, V. J. et al. Systematic study of the glutathione (GSH) reactivity of N-arylacrylamides: 1. Effects of aryl substitution. J. Med. Chem. 58, 9171-9178 (2015).
- Jackson, P. A., Widen, J. C., Harki, D. A. & Brummond, K. M. Covalent modifiers: a chemical perspective on the reactivity of α,β-unsaturated carbonyls with thiols via hetero-Michael addition reactions. J. Med. Chem. 60, 839-885 (2017).
- Nichols, R. J. et al. RAS nucleotide cycling underlies the SHP2 phosphatase dependence of mutant BRAF-, NF1- and RAS-driven cancers, Nat. Cell Biol. 20, 1064-1073 (2018).
- Robert, C. et al. Improved overall survival in melanoma with combined dabrafenib and trametinib, N. Engl. J. Med. 372, 30-39 (2015)
- Saiki, A. Y. et al. MDM2 antagonists synergize broadly and robustly with compounds targeting fundamental oncogenic signaling pathways. Oncotarget 5, 2030-2043 (2014).
- Ebert, P. J. R. et al. MAP kinase inhibition promotes T cell and anti-tumor activity in combination with PD-L1 checkpoint blockade. Immunity 44, 609-621 (2016).
- Selby, M. J. et al. Preclinical development of ipilimumab and nivolumab combination immunotherapy: mouse tumor models, in vitro functional studies, and cynomolgus macaque toxicology. PLoS ONE 11, e0161779 (2016).
- Mosely, S. I. et al. Rational selection of syngeneic preclinical tumor models for immunotherapeutic drug discovery. Cancer Immunol. Res. 5, 29-41 (2017).
- 27. Spranger, S., Dai, D., Horton, B. & Gajewski, T. F. Tumor-residing Batf3 dendritic cells are required for effector T cell trafficking and adoptive T cell therapy. Cancer Cell 31, 711-723 (2017)
- Gao, Q. et al. Cancer-cell-secreted CXCL11 promoted CD8<sup>+</sup>T cells infiltration through docetaxel-induced-release of HMGB1 in NSCLC, J. Immunother, Cancer 7, 42 (2019).
- Lee, J. W. et al. The combination of MEK inhibitor with immunomodulatory antibodies targeting programmed death 1 and programmed death ligand 1 results in prolonged survival in Kras/p53-driven lung cancer. Journal Thorac. Oncol. 14, 1046-1060 (2019).
- Liu, L. et al. The BRAF and MEK inhibitors Dabrafenib and Trametinib: Effects on Immune Function and in Combination with Immunomodulatory antibodies targeting PD-1, PD-L1, and CTLA-4. Clin. Cancer Res. 21, 1639-1651 (2015).
- Kordiak, J. et al. Intratumor heterogeneity and tissue distribution of KRAS mutation in nonsmall cell lung cancer; implications for detection of mutated KRAS oncogene in exhaled breath condensate. J. Cancer Res. Clin. Oncol. 145, 241-251 (2019).
- Lamy, A. et al. Metastatic colorectal cancer KRAS genotyping in routine practice: results and pitfalls, Mod. Pathol. 24, 1090-1100 (2011).
- Richman, S. D. et al. Intra-tumoral heterogeneity of KRAS and BRAF mutation status in patients with advanced colorectal cancer (aCRC) and cost-effectiveness of multiple sample testing. Anal. Cell. Pathol. 34, 61-66 (2011).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Acknowledgements We thank N. Moua Vang, P. Achanta, J. Estrada, P. Mitchell, T. Tsuruda, D. Mohl, C. Liu, J. Lofgren, R. Shimanovich, P. Agarwal, R. S. Foti, Y. B. Yu, J. Yadav, M. Singh, J. Nam, C. Wang, R. Pham, W. Rufai, T. McElroy, S. Tiso, M. Farley, J. Ngang, D. Wu, R. Dawson, J. Reidy, J. Egen, R. Kendall, P. J. Beltran, M. Eschenberg, S. Caenepeel, P. Hughes, A. Coxon, F. Martin, P. K. Morrow, S. Agrawal, D. Nagorsen and G. Friberg for their support and contributions; all of the patients who participated in the AMG 510 first-in-human clinical trial; and the staff of Crystallographic Consulting and the Advanced Light Source at beamline 5.0.1 for their data collection support. The Berkeley Center for Structural Biology is supported in part by the National Institutes of Health, National Institute of General Medical Sciences and the Howard Hughes Medical Institute. The Advanced Light Source is supported by the Director, Office of Science, Office of Basic Energy Sciences, of the US Department of Energy under contract no. DE-ACO2-05CH11231.

Author contributions B.A.L. oversaw the design and synthesis of compounds. C.M. solved the crystal structure of AMG 510. J.D.M. and T.A. designed the nucleotide-exchange assay and mass spectrometry experiment with recombinant KRAS, and T.S.M. and A.Y.S. developed the assays. L.P.V. and C.G.K. oversaw the bioanalytical assessment of AMG 510. J.-R.S., T.H., K.C., K.R., A.Y.S. and T.O. executed and analysed in vivo studies. L.P.V., X.Z. and R.O. developed methods, and N.K. quantified KRAS(G12C)–AMG 510 covalent adducts in cells and tumour samples. L.P.V., R.O. and X.Z. developed and executed the SILAC study to determine the half-life of KRAS(G12C). J.R.L., A.Y.S., T.A., W.O. and V.J.C. designed, and A.Y.S., T.S.M., A.S.R. and K.G. executed in vitro experiments. J.W. generated the immunohistochemical data. D.B. designed and analysed proteomic experiments. B.E.H. provided clinical pharmacokinetic data. J.C., K.R. and K.C. designed the in vivo experiments. H.H. oversaw the clinical development. M.G.F., B.H.O., T.J.P., G.S.F., J.D., J.K., R.G. and D.S.H. were investigators for the AMG 510 clinical trial. J.C., J.R.L., V.J.C., A.Y.S. and K.R. wrote the paper with contributions from all authors.

Competing interests J.C., K.R., A.Y.S., C.M., K.C., D.B., K.G., T.H., C.G.K., N.K., B.A.L., J.W., A.S.R., R.S.M., R.O., T.O., J.-R.S., X.Z., J.D.M., L.P.V., B.E.H., W.O., H.H., T.A., V.J.C. and J.R.L. were employees and stock holders of Amgen at the time of data collection. M.G.F. has received grant and research support from AstraZeneca, Amgen and Novartis; served as a consultant for Array BioPharma, Amgen and Seattle Genetics; and has been part of the speakers bureau for Amgen. B.H.O. has received honoraria from Amgen. T.J.P. has received grants and research support from Amgen. G.S.F. is employed by HealthONE and the Sarah Cannon Research

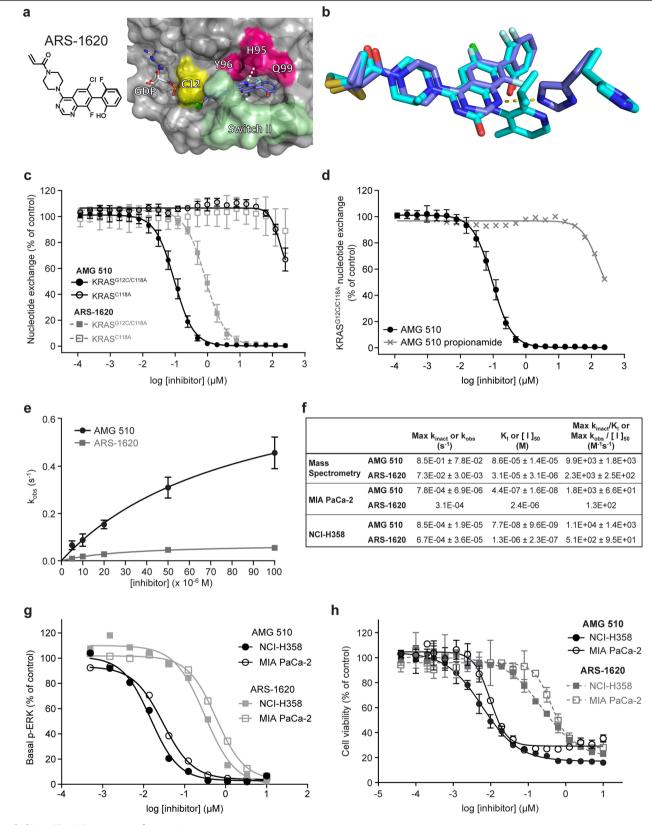
Institute; has served in a consulting or advisory capacity for EMD Serono and Fujifilm; has received research funding from 3-V Biosciences, Abbvie, ADC Therapeutics, Aileron, American Society of Clinical Oncology, Amgen, ARMO, AstraZeneca, BeiGene, Bioatla, Biothera, Celldex, Celgene, Ciclomed, Curegenix, Curis, DelMar, eFFECTOR, Eli Lilly, EMD Serono, Exelixis, Fujifilm, Genmab, GlaxoSmithKline, Hutchison MediPharma, Ignyta, Incyte, Jacobio, Jounce, Kolltan, Loxo, MedImmune, Millennium, Merck, Mirna Therapeutics, the National Institutes of Health, Novartis, OncoMed, Oncothyreon, Precision Oncology, Regeneron, Rgenix, Ribon, Strategia, Syndax, Taiho, Takeda, Tarveda, Tesaro, Tocagen, Turning Point Therapeutics, the UT MD Anderson Cancer Center and Vegenics; receives royalties from Wolters Kluwer; and has also received travel, accommodation and related expenses from Bristol-Myers Squibb, EMD Serono, Fujifilm, Millennium and the Sarah Cannon Research Institute. J.D. has served in a consulting or advisory capacity for Amgen, BeiGene, Bionomics, Eisai, Eli Lilly and Novartis; and received research funding from Bionomics, GlaxoSmithKline, Novartis and Roche. J.K. has received travel, accommodation and expenses from Bristol-Myers Squibb, MSD and Zucero Therapeutics. R.G. has served in a consulting or advisory capacity for AbbVie, Adaptimmune, AstraZeneca, Celgene, Ignyta, Inivata, Merck, Nektar, Pfizer and Roche, D.S.H. has stock and other ownership interests in Molecular Match. OncoResponse and Presagia: has served in a consulting or advisory capacity for Alpha Insights, Axiom, Adaptimmune, Baxter, Bayer, Genentech, GLG, Group H, Guidepoint Global, Infinity, Janssen, Merrimack, Medscape, Numab, Pfizer, Seattle Genetics, Takeda and Trieza Therapeutics: has received research and/or grant funding from AbbVie, Adaptimmune, Amgen, AstraZeneca, Bayer, BMS, Daiichi-Sankyo, Eisai, Fate Therapeutics, Genentech, Genmab, Ignyta, Infinity, Kite, Kyowa, Eli Lilly, LOXO, Merck, MedImmune, Mirati, Mirna Therapeutics, Molecular Templates, Mologen, NCI-CTEP, Novartis, Pfizer, Seattle Genetics and Takeda; and has received travel, accommodation and expenses from Genmab, Loxo Oncology, ASCO, AACR, SITC and Mirna Therapeutics.

#### Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1694-1

**Correspondence and requests for materials** should be addressed to J.C. or J.R.L. **Peer review information** *Nature* thanks Rene Bernards and the other, anonymous, reviewer(s)

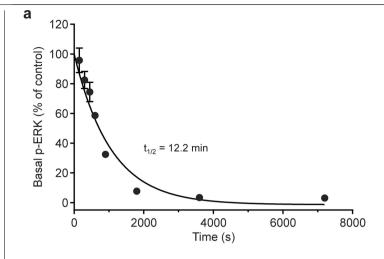
for their contribution to the peer review of this work. **Reprints and permissions information** is available at http://www.nature.com/reprints.

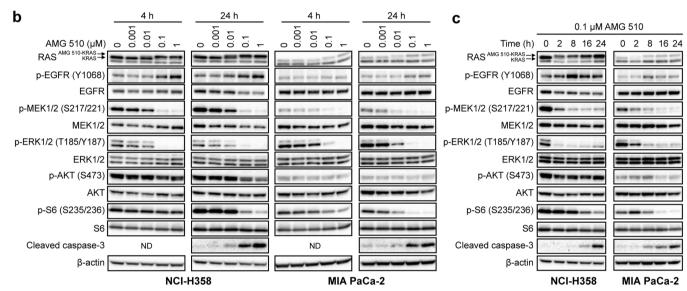


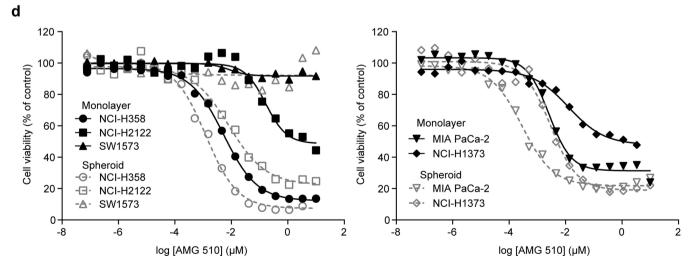
 $\textbf{Extended Data Fig. 1} | See \ next \ page \ for \ caption.$ 

**Extended Data Fig. 1**| **Enhanced binding of AMG 510 to KRAS(G12C) results in improved properties. a**, X-ray co-crystal structure of KRAS(G12C/C51S/C80L/C118S) bound to GDP and ARS-1620 (PDB: 5V9U). **b**, Overlay of ARS-1620 and AMG 510. The right side shows different orientations of His95 (H95) depending on the ligand. **c**, Biochemical activity of AMG 510 and ARS-1620 in a nucleotide-exchange assay with purified KRAS(G12C/C118A) or KRAS(C118A) protein. Data are mean  $\pm$  s.d., n = 4 replicates. The wild-type cysteine at position 118 was changed to alanine to avoid reactivity with non-Cys12 cysteines. **d**, Biochemical activity of AMG 510 and its non-reactive propionamide analogue

in a nucleotide-exchange assay with purified KRAS(G12C/C118A); propionamide, mean of n=2 replicates.  $\mathbf{e}$ , Kinetic properties of AMG 510 and ARS-1620 as determined by mass spectrometry.  $\mathbf{f}$ , Calculated maximal reaction rates ( $k_{\text{inact}}$  or  $k_{\text{obs}}$ ) and the concentrations that achieve a half-maximal rate ( $K_1$  or  $[I]_{50}$ ) of AMG 510 and ARS-1620.  $\mathbf{e}$ ,  $\mathbf{f}$ ,  $k_{\text{obs}}$ ,  $K_1$ ,  $[I]_{50}$  and standard error of the curve were determined from nonlinear curve fitting of experimental values.  $\mathbf{g}$ ,  $\mathbf{h}$ , Inhibition of p-ERK after a 2-h treatment ( $\mathbf{g}$ ; mean, n=2 replicates) and effects on cell viability after 72-h treatment ( $\mathbf{h}$ ; mean  $\pm$  s.d., n=3 replicates) with AMG 510 or ARS-1620.

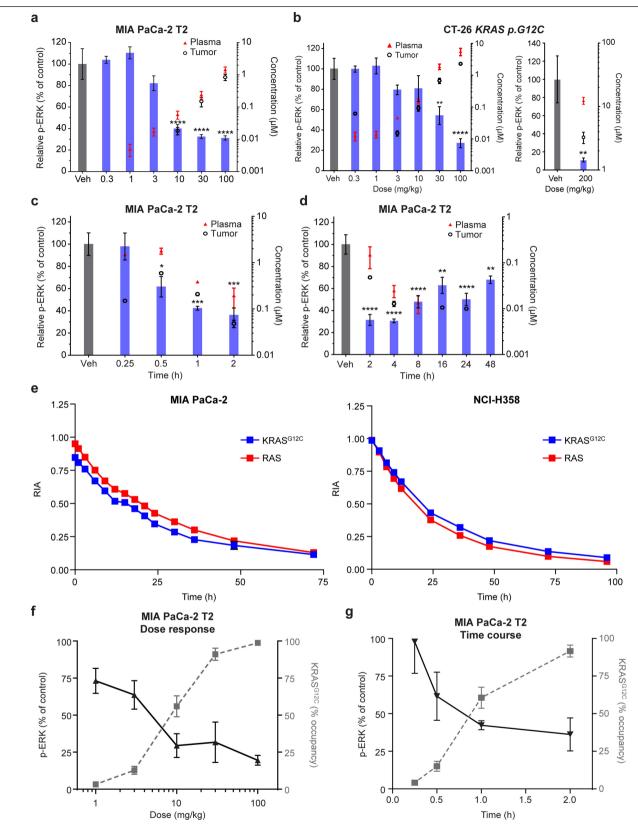






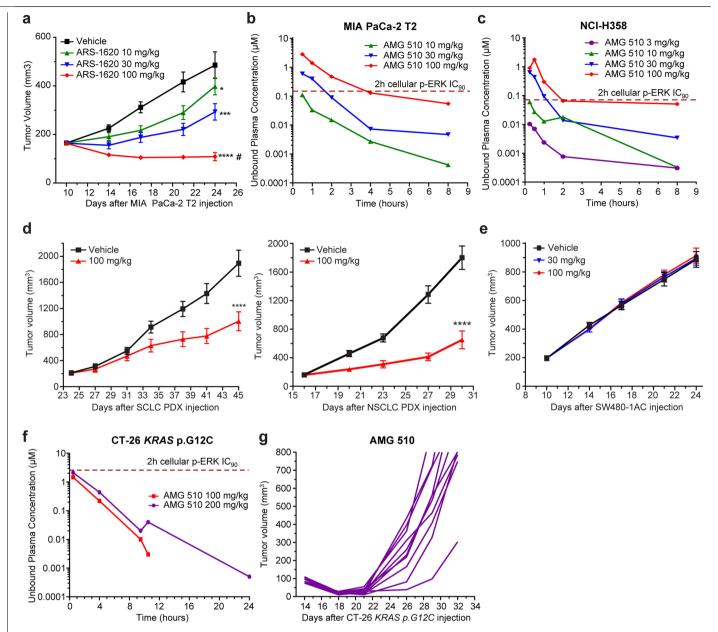
Extended Data Fig. 2 | AMG 510 inhibits KRAS(G12C) signalling and impairs viability. a, Inhibition of p-ERK with RMC-4550 in NCI-H358 cells. Data are mean  $\pm$  s.d., n = 3 replicates. b, c, Effect on cellular signalling in NCI-H358 or MIA PaCa-2 after 4- or 24-h treatment with a serial titration of AMG 510 (b) or treatment with 0.1  $\mu$ M AMG 510 at time points for up to 24 h (c). Top arrow, AMG

510–KRAS (G12C) covalent adduct; bottom arrow, KRAS. Data are from a single experiment (Supplementary Fig. 1). **d**, Effect of 72-h treatment with AMG 510 on cell viability in adherent monolayer or spheroid culture conditions (mean, n=2 replicates).



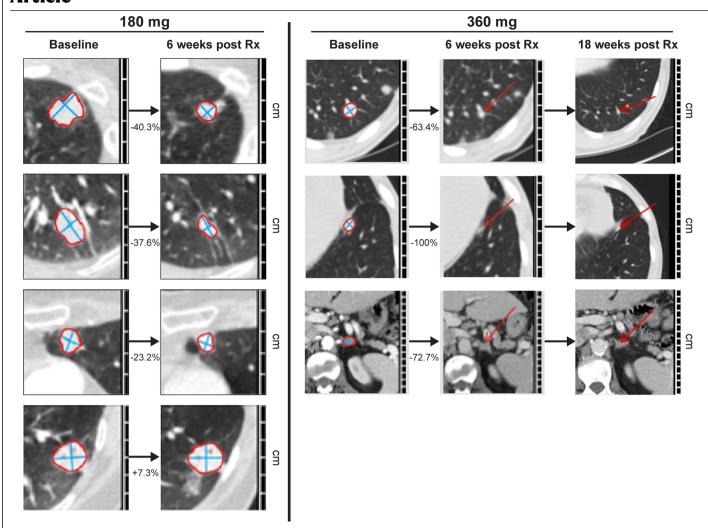
**Extended Data Fig. 3** | **AMG 510 covalently modifies KRAS(G12C) in tumours and inhibits signalling in vivo.**  $\mathbf{a}$ - $\mathbf{d}$ , Mice bearing MIA PaCa-2 T2 ( $\mathbf{a}$ ,  $\mathbf{c}$ ,  $\mathbf{d}$ ) or CT-26  $KRAS^{Ca2C}(\mathbf{b})$  tumours were treated orally with a single dose of vehicle (black bars) or with the indicated doses of AMG 510 (bluebars). Tumours were collected 2 h later ( $\mathbf{a}$ ,  $\mathbf{b}$ ) or over time as indicated ( $\mathbf{c}$ ,  $\mathbf{d}$ ) and levels of p-ERK were measured. AMG 510 concentrations in plasma (red triangles) or tumours (black open circles). Data are mean  $\pm$  s.e.m., n = 3 mice per group; \*\*\*\*P< 0.0001, \*\*\*\*P< 0.001,

\*\*P< 0.01 compared with vehicle; one-way ANOVA followed by Dunnett's multiple-comparison test. **e**, Half-life determination of KRAS(G12C) in MIA PaCa-2 and NCI-H358 cells by SILAC. Data are mean  $\pm$  s.d., n = 3 replicates. **f**, **g**, AMG 510 treatment results in covalent modification of KRAS(G12C) that inversely correlates with p-ERK inhibition in MIA PaCa-2T2 tumours. Data are mean  $\pm$  s.d., n = 3 mice per group.



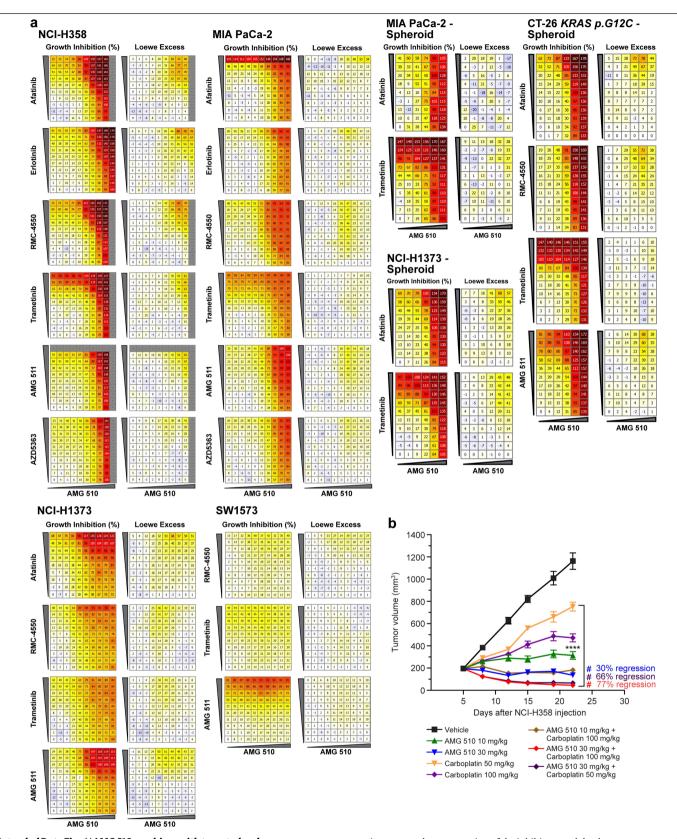
Extended Data Fig. 4 | AMG 510 inhibits tumour growth of patient-derived xenografts, and exposure to AMG 510 at or above cellular IC $_{90}$  drives regression of xenografts. a, Mice bearing MIA PaCa-2T2 tumours were treated with ARS-1620 at the indicated doses. Data are mean  $\pm$  s.e.m., n = 10 mice per group; \*\*\*\*P<0.0001, \*\*\*P<0.001, \*P<0.05 compared with vehicle; repeated-measures ANOVA followed by Dunnett's multiple-comparison test. \*P<0.05 regression by two-sided Student's P-test. P-

treatment on tumour growth in  $KRAS^{G12C}$  small-cell lung cancer (SCLC) and non-small-cell lung cancer (NSCLC) PDX models (**d**) or a SW480-1AC xenograft model (**e**). Data are mean  $\pm$  s.e.m., n = 10 mice per group. \*\*\*\*P< 0.0001 compared with vehicle; repeated-measures ANOVA followed by Dunnett's multiple-comparison test. **f**, Plasma levels of AMG 510 from CT-26  $KRAS^{G12C}$  tumour model. The dotted line represents the p-ERK IC  $_{90}$  values in cells after treatment with AMG 510 for 2 h. **g**, Individual CT-26  $KRAS^{G12C}$  tumour plots of BALB/c nude mice (n = 10) treated with AMG 510 (200 mg kg $^{-1}$ ).



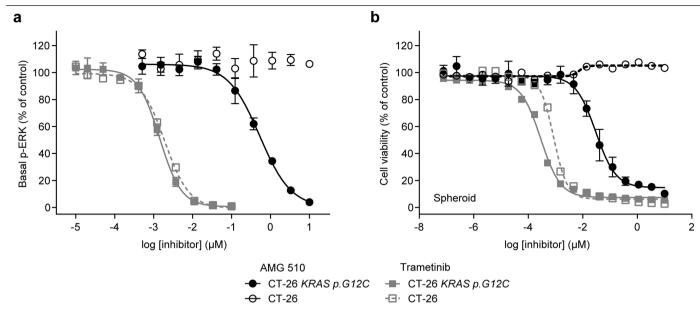
Extended Data Fig. 5 | Clinical activity of AMG 510 in patients with lung cancer in a first-in-human dose-escalation study. Computed tomography scans of two patients with  $KRAS^{GIZC}$  lung carcinoma who were treated with AMG 510. Additional representative pre-treatment (baseline) and post-treatment ( $R_x$ ) scans of patients described in Fig. 3 (left, 180 mg; right, 360 mg). Lesions are

denoted by red outline or red arrows. Left images show, from top to bottom, the lung upper left lobe, lung lower left lobe, lung upper left lobe and lung upper left lobe. Right images show, from top to bottom, the lung lower left lobe, lung lower left lobe and adrenal gland. Lesions in the 18-week scans of the patient who was treated with 360 mg AMG 510 were considered too small to accurately measure.



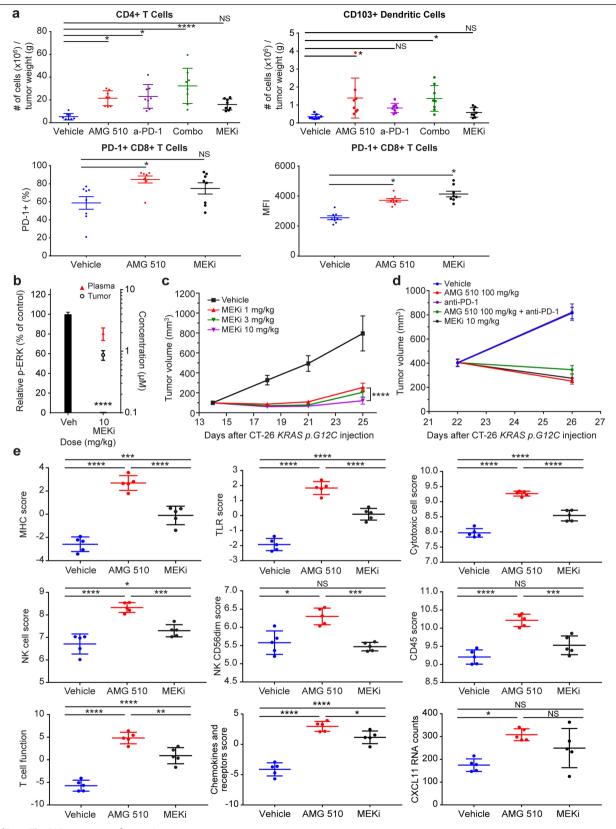
 $\label{lem:extended} \textbf{Data Fig. 6} | \textbf{AMG 510 combines with targeted and chemotherapeutic agents, resulting in the synergistic killing of tumour cells and enhanced anti-tumour activity. a, Growth inhibition matrices and Loewe additivity excess of AMG 510 added in combination with targeted agents to the indicated cell line, with darker colours denoting greater cell killing (growth inhibition) and stronger synergistic interactions (Loewe excess). The$ 

maximum tested concentration of the inhibitors and the dose range covered by the matrices for each combination are listed in Supplementary Table 3. **b**, AMG 510 in combination with carboplatin in NCI-H358 tumour xenografts. Data are mean  $\pm$  s.e.m., n = 10 mice per group; \*\*\*\*\*P < 0.0001 compared with vehicle; repeated-measures ANOVA followed by Dunnett's multiple comparison test; \*P < 0.001 regression by two-sided Student's t-test.



**Extended Data Fig. 7** | **AMG 510 inhibits KRAS(G12C) signalling and viability of syngeneic CT-26**  $KRAS^{G12C}$  **cells. a, b,** Cellular activity of AMG 510 and the MEK inhibitor trametinib in CT-26  $KRAS^{G12C}$  and parental CT-26 cell lines as measured by the inhibition of p-ERK after a 2-h treatment (a; mean  $\pm$  s.d., n = 3 replicates;

except trametinib in CT-26, mean of n=2 replicates) and the effects on cell viability after 72-h treatment in spheroid culture (**b**; AMG 510 in CT-26 *KRAS* <sup>G12C</sup>, mean  $\pm$  s.d., n=3 replicates; all others, mean of n=2 replicates).

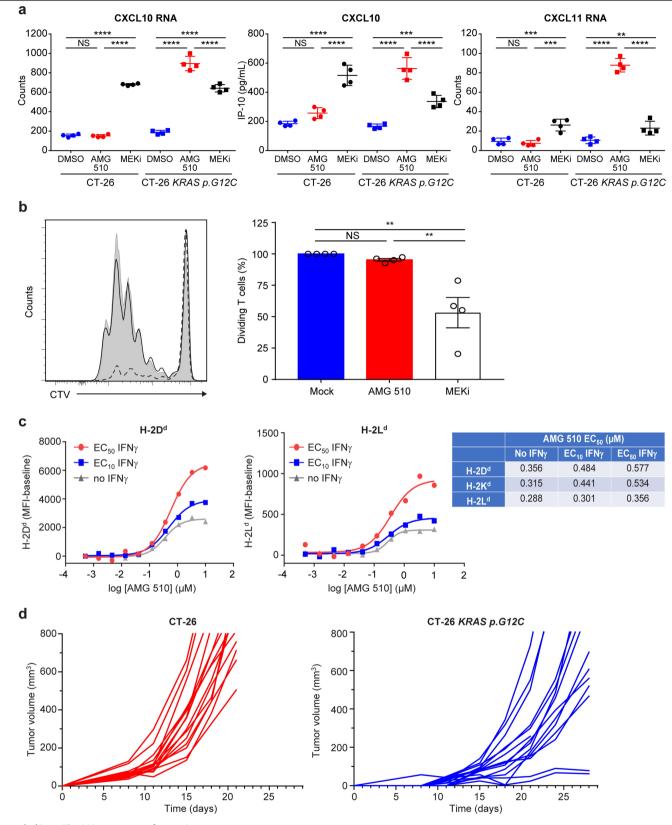


 $\textbf{Extended Data Fig. 8} \, | \, \textbf{See next page for caption.}$ 

# $\label{eq:continuous} \textbf{Extended Data Fig. 8} \ | \ \textbf{AMG 510 treatment induces a pro-inflammatory tumour microenvironment. a}, \ \textbf{CT-26} \ \textit{KRAS}^{\textit{G12C}} \ tumours \ were$

immunophenotyped by flow cytometry. Data are mean  $\pm$  s.d., n=8 mice per group; \*\*\*\*P<0.0001, \*P<0.05; NS, not significant; one-way ANOVA followed by Tukey's multiple-comparison test. MEKi, MEK inhibitor. **b**, Mice bearing CT-26 KRAS<sup>GIZC</sup> tumours were treated or ally with a single dose of vehicle (black bar) or with the indicated dose of MEK inhibitor (blue bar). Tumours were collected 2 h later and levels of p-ERK were measured. MEK inhibitor concentrations in plasma (red triangle) or tumours (black open circle). Data are mean  $\pm$  s.e.m., n=3 mice per group; \*\*\*\*P<0.0001 compared with vehicle; one-way ANOVA followed by

Dunnett's multiple-comparison test. **c**, Mice bearing CT-26  $KRAS^{GI2C}$  tumours were treated with MEK inhibitor at the indicated doses. Data are mean  $\pm$  s.e.m., n=8 mice per group; \*\*\*\*P<0.0001 compared with vehicle; repeated-measures ANOVA followed by Dunnett's multiple comparison test. **d**, Tumour volumes from the immunophenotyping study (**a**) of CT-26  $KRAS^{GI2C}$  tumour-bearing mice treated over 4 days. n=8 mice per group. **e**, RNA was isolated from CT-26  $KRAS^{GI2C}$  tumours. n=5 mice per group. Gene expression and scores were calculated by nSolver v.4.0. Data are mean  $\pm$  s.d.; \*\*\*\*P<0.0001, \*\*P<0.005; NS, not significant; one-way ANOVA followed by Tukey's test.



 $\textbf{Extended Data Fig. 9} \, | \, \textbf{See next page for caption}.$ 

**Extended Data Fig. 9** | **AMG 510 induces expression of chemokines and MHC class I antigens in CT-26**  $KRAS^{GIZC}$  **cells. a**, Quantification of Cxcl10 or Cxcl11 transcripts, as well as secreted CXCL10 (IP-10) protein, after 24-h treatment of parental CT-26 or CT-26  $KRAS^{GIZC}$  cells with AMG 510 or MEK inhibitor. Data are mean  $\pm$  s.d., n = 4 replicates; \*\*\*\*P < 0.0001, \*\*\*P < 0.001, \*\*P < 0.01; NS, not significant; two-way ANOVA followed by Tukey's multiple-comparison test. **b**, Ova-pulsed bone-marrow-derived dendritic cells and CellTrace Violet (CTV)-labelled OT-I CD8\* T cells co-cultured with AMG 510 or MEK inhibitor. T cell proliferation was assessed by measuring CTV dilution in T cells. Left, T cells

treated with mock (shaded), AMG 510 (solid line) or MEK inhibitor (dashed line) from a representative experiment. Right, data from four independent experiments were pooled and show the frequency of dividing T cells relative to mock treatment. Data are mean  $\pm$  s.e.m.; \*\*P<0.01; NS, not significant; one-way ANOVA followed by Tukey's multiple-comparison test. **c**, Cell surface expression of MHC class I antigens (H-2D<sup>d</sup> and H-2L<sup>d</sup>) on CT-26 *KRAS* <sup>G12C</sup> cells after 24-h treatment with AMG 510 with or without IFNy as measured by flow cytometry. **d**, Growth curves of either CT-26 or CT-26 *KRAS* <sup>G12C</sup> tumours in BALB/c mice

## Extended Data Table 1 | Data collection and refinement statistics for AMG 510-KRAS(G12C) complex

	KRAS <sup>G12C/C51S/C80L/C118S</sup>
	AMG 510 ( <b>60IM</b> )
Data collection	
Space group	P 21 21 21
Cell dimensions	
a, b, c (Å)	40.87, 58.42, 65.89
$\alpha, \beta, \gamma$ (°)	90, 90, 90
Resolution (Å)	30.0-1.65 (1.71-1.65)
$R_{ m sym}$	0.162 (0.521)
$I/\sigma I$	6.9 (2.5)
Completeness (%)	97.0 (96.3)
Redundancy	4.4 (4.2)
D.C., 4	
Refinement	20.00 1.65
Resolution (Å)	30.00 - 1.65
No. reflections	18077
$R_{ m work}$ / $R_{ m free}$	0.1809 / 0.2152
No. atoms	1613
Protein	1336
Ligand/ion	70
Water	207
B-factors	24.8
Protein	24.3
Ligand/ion	24.1
Water	34.1
R.m.s. deviations	
Bond lengths (Å)	0.005
Bond angles (°)	1.08

 $One crystal \ dataset \ was collected for the \ X-ray \ co-crystal \ structure \ of the \ AMG 510-KRAS (G12C) \ covalent \ complex. \ Values \ in \ parentheses \ are for the \ highest-resolution \ shell.$ 



Last updated by author(s): Sep 3, 2019

# **Reporting Summary**

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see <u>Authors & Referees</u> and the <u>Editorial Policy Checklist</u>.

<u> </u>				
St	ลา	715	ŤΙ	CS

For	all st	atistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.
n/a	Cor	nfirmed
	$\boxtimes$	The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
	$\boxtimes$	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
	$\boxtimes$	The statistical test(s) used AND whether they are one- or two-sided  Only common tests should be described solely by name; describe more complex techniques in the Methods section.
X		A description of all covariates tested
	$\boxtimes$	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
	$\boxtimes$	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
$\boxtimes$		For null hypothesis testing, the test statistic (e.g. $F$ , $t$ , $r$ ) with confidence intervals, effect sizes, degrees of freedom and $P$ value noted Give $P$ values as exact values whenever suitable.
$\boxtimes$		For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
$\times$		For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
$\boxtimes$		Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated
		Our web collection on statistics for biologists contains articles on many of the points above

## Software and code

Policy information about <u>availability of computer code</u>

Data collection

B3 v3 (x-ray crystallography)

EnVision Manager v1.13-1.14 (nucleotide exchange, viability, combination assays)

Discovery Workbench v4.0 (p-ERK assays) ImageLab v4.1 and v5.2.1 (immunoblotting)

Agilent RapidFire v4.0, MassHunter Workstation B.05.01 (mass spectrometry kinetic assay)

nSolver v4.0 (NanoString)

StudyDirector v3.1 (in vivo studies) BD FACSDiva v8.0.1(flow cytometry) Immunospot v2.6.1 (ELIspot)

Analyst v1.6 (AMG 510-KRAS G12C conjugate detection, SILAC)

Data analysis

CCP4 Program Suite v6.4.0, HKL2000 v717, MolRep v11.2.08, Refmac5 v5.8.0073, Coot v0.7.2, PRODRG v050106.0517 (x-ray crystallography)

Microsoft Excel for Office 365 (nucleotide exchange, viability, p-ERK, kinetic, ELIspot, cysteine proteomics, flow cytometry)

GraphPad Prism v7.04 (nucleotide exchange, viability, p-ERK, kinetic, in vivo efficacy/survival, flow cytometry, NanoString, ELIspot, SILAC)

MassHunter Qualitative Analysis B.07.00 (mass spectrometry kinetic assay)

Chalice Analyzer v1.5.0.71 (combination synergy scores)

SEQUEST (cysteine proteomics) nSolver v4.0 (NanoString) Mathematica v11.3 (SILAC)

Immunospot v2.6.1 (ELIspot)

BD FACSDiva v8.0.1, FlowJo software v10 (flow cytometry)

Biostatistical Analysis R Shiny application v1.0.5 (in vivo PKPD studies)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The majority of data generated or analyzed during this study are included in this published article or available at the source data links. X-ray crystallographic coordinates and structure factor files have been deposited in the Protein Data Bank (PDB ID code: 60IM). Other data that support the findings of this study are available from the corresponding authors. Qualified researchers may request data from Amgen clinical studies. Complete details are available here: http:// www.amgen.com/datasharing.

Field-spe	cific reporting		
Please select the or	ase select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.		
Life sciences	Behavioural & social sciences Ecological, evolutionary & environmental sciences		
For a reference copy of the	ne document with all sections, see <u>nature.com/documents/nr-reporting-summary-flat.pdf</u>		
Life scien	ices study design		
All studies must disc	close on these points even when the disclosure is negative.		
Sample size	For in vivo PKPD studies, n=3/group were used. For efficacy studies, n=8-10 mice per group were used. Animal numbers for in vivo studies were selected using power analysis alpha 0.05 and 80% power such that a minimum change of 32-49% could be detected on the observed data scale. No sample size calculation was performed for in vitro studies.		
Data exclusions	No data were excluded.		
Replication	In vitro experiments were repeated the indicated number of times, with the exception of immunoblot experiments which were performed once. Synergy scores were determined from the aggregate of two 10x10 matrices for adherent monolayer combinations, but only one 6x10 matrix for spheroid combinations. In vivo PKPD dose response studies (MIA PaCa-2 T2, CT-26 KRAS p.G12C) were repeated with similar results at least twice. The combination of AMG 510 with anti-PD-1 in CT-26 KRAS p.G12C, as well as the tumor growth measurements of untreated CT-26 parental and CT-26 KRAS p.G12C tumors, were repeated twice with similar results. All other in vivo studies were performed once. Independent repeats and sample sizes, as well as statistical analyses and significance levels, are also indicated in the Figure legends or in the Statistics and Reproducibility section.		
	Statistics and Reproducibility Section.		
Randomization	Sample randomization is not relevant to the in vitro studies presented. For in vivo studies, animals were evenly distributed such that each group had a similar mean and SEM at the start of the study.		
Blinding	Treatment groups for the in vivo combination studies were blinded to the investigator.		

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems		Methods
n/a	Involved in the study	n/a Involved in the study
	Antibodies	ChIP-seq
	Eukaryotic cell lines	Flow cytometry
$\times$	Palaeontology	MRI-based neuroimaging
	Animals and other organisms	
	Human research participants	
	Clinical data	

## **Antibodies**

Antibodies used

Immunoblot: all antibodies were used at 1:1,000 dilution unless otherwise indicated. phospho-EGF Receptor (Tyr1068) (D7A5) XP® Rabbit mAb Cell Signaling #3777; Lot 13 EGF Receptor (D38B1) XP® Rabbit mAb Cell Signaling #4267; Lot 11

Phospho-MEK1/2 (Ser217/221) Antibody Cell Signaling #9121; Lot 44

MEK1/2 Antibody Cell Signaling #9122; Lot 14

Phospho-S6 Ribosomal Protein (Ser235/236) (D57.2.2E) XP® Rabbit mAb Cell Signaling #4858; Lot 16

S6 Ribosomal Protein (5G10) Rabbit mAb Cell Signaling #2217; Lot 7

Phospho-Akt (Ser473) (D9E) XP® Rabbit mAb Cell Signaling #4060; Lot 19

Akt Antibody Cell Signaling #9272, Lot 27

Phospho-ERK1/ERK2 (Thr185, Tyr187) Polyclonal Antibody ThermoFisher #44-680G; Lot SB248818

p44/42 MAPK (Erk1/2) Antibody Cell Signaling #9102; Lot 26

Anti-Ras antibody [EPR3255] Abcam #ab108602; Lot GR117071-23

Cleaved Caspase-3 (Asp175) Antibody Cell Signaling #9661; Lot 45

Anti-β-Actin-Peroxidase Mouse mAb AC-15 Sigma #A3854; Lot 026M4820V; 1:20,000

Donkey Anti-rabbit IgG HRP GE Healthcare #NA934V; Lot 9677977; 1:5,000

Flow cytometry: all antibodies were used at 1:100 unless otherwise indicated.

PE Mouse anti-mouse H-2Dd (34-2-12, Biolegend #110608, Lot B256526)

PE Mouse anti-mouse H-2Kd (SF1-1.1, Biolegend #116608, Lot B244820)

PE Mouse anti-mouse H-2Ld/H-2Db (28-14-8, Biolegend #114507, Lot B240332)

BUV737 rat anti-mouse CD4 (RM4-5, BD Biosciences #564933, Lot 8164630)

BV421 rat anti-mouse CD8a (53-6.7, BD Biosciences #563898, Lot 7201962)

BUV737 rat anti-CD11b (M1/70, BD Biosciences #564443, Lot 7338572)

BV786 mouse anti-mouse CD45.2 (104, BD Biosciences #563686, Lot 8235903)

APC-H7 rat anti-mouse Ly-6G (1A8, BD Biosciences #565369, Lot 8121728)

BV711 hamster anti-mouse TCR B chain (H57-597, BD Biosciences #563135, Lot 7054698)

FITC rat anti-mouse CD24 (M1/69, ThermoFisher #11-0242-81, Lot 1937898)

APC hamster anti-mouse CD103 (2E7, ThermoFisher #17-1031-80, Lot 17-1031-80) PE hamster anti-mouse CD279 (PD-1) (J43, ThermoFisher#12-9985-82, Lot 4329622)

APC/Cy7 rat anti-mouse CD90.2 (30-H12, Biolegend #105328, Lot B241601)

BV650 rat anti-mouse F4/80 (BM8, Biolegend #123149, Lot B256505)

BV711 rat anti-mouse Ly-6C (HK1.4, Biolegend #128037, Lot B247973)

BV510 rat anti-mouse I-A/I-E (MHCII) (M5/114.15.2, Biolegend #107635, Lot B263357)

APC rat anti-mouse CD8a (53-6.7, eBioscience #17-0081-82, Lot E07056-1635)

rat anti-mouse CD16/CD32 (2.4G2, BD Biosciences #553142, Lot 4198965)

FITC hamster anti-mouse TCR B chain (H57-597, BD Biosciences #553171, Lot 8351664)

#### Immunohistochemistry:

Rat anti-human CD3 (mouse CD3 cross-reactive) (CD3-12, Bio-Rad #MCA1477, Lot 7708); 1:1,000

Rabbit anti-mouse CD8a (D4W2Z, Cell Signaling Technology #98941, Lot 0712017); 1:500

Rabbit anti-human Ki67 (mouse Ki67 cross-reactive) (SP6, Sigma-Aldrich #275R-1, Lot 45305); 1:500

Rat IgG isotype negative control (Jackson Immunoresearch Labs #012-000-003, Lot 68714); 2 mcg/mL Rabbit IgG isotype negative control (Jackson Immunoresearch Labs #011-000-003, Lot 132409); 2 mcg/mL

HRP-anti-rat-lgG (Biocare Medical #BRR4016L, Lot 100317); undiluted

HRP-anti-rabbit-IgG (Dako #K4003, Lot 10147964); undiluted

Validation

All antibodies were validated by the manufacturer. Please refer to the manufacturers' websites with the catalog information listed above.

## Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)

The following cell lines were purchased from American Type Culture Collection (ATCC): MIA PaCa-2, NCI-H1373, NCI-H2030, NCI-H2122, SW1463, SW1573, UM-UC-3, Calu-1, NCI-H1792, NCI-H23, NCI-H358, SW837, AsPC-1, A-427, LS 174T, SW480, A549, NCI-H1355, HCC-827, COLO-205. KM12 and NCI-H3122 were obtained from the Amgen internal cell bank, originally sourced from the National Cancer Institute.

MIA PaCa-2 T2 and SW480-1AC cells were generated by passaging MIA PaCa-2 and SW480 cells, respectively, in mice.

CT-26 KRAS p.G12C cells were generated from the murine CT-26 colorectal line (ATCC) using CRISPR technology to replace both KRAS p.G12D alleles with p.G12C (ThermoFisher Scientific).

Authentication

Cell lines were authenticated by short tandem repeat (STR) profiling or were used immediately after purchase from ATCC.

Mycoplasma contamination

All cell lines used for in vivo studies were confirmed to be negative for mycoplasma contamination.

Commonly misidentified lines (See ICLAC register)

No commonly misidentified cell lines were used.

## Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals

BALB/c or athymic nude mice, all female, all 6-12 weeks of age.

Wild animals Studies did not involve wild animals

Studies did not involve samples collected in the field. Field-collected samples

Ethics oversight All animal experimental protocols were approved by the Amgen Animal Care and Use Committee and were conducted in accordance with the guidelines set by the Association for Assessment and Accreditation of Laboratory Animal Care.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Human research participants

Population characteristics

Policy information about studies involving human research participants

See clinicaltrials.gov NCT03600883.

Key inclusion criteria: age ≥18; documented locally-advanced or metastatic KRASG12C; measurable or evaluable disease; ECOG ≤2; life expectancy >3 months (mo). Key exclusion criteria: active brain metastases; myocardial infarction within 6 mo.

Recruitment Patients were recruited at clinical study sites based on the presence of the KRAS p.G12C mutation in their tumor by standard genotype testing.

institutional review boards (IRB)/independent ethics committees (IEC) of all clinical study sites.

Clinical trial NCT03600883 was conducted in compliance with all relevant ethical regulations. The protocol was approved by the Ethics oversight

Note that full information on the approval of the study protocol must also be provided in the manuscript.

#### Clinical data

Data collection

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

NCT03600883 Clinical trial registration

Study protocol Study is ongoing; clinical trial information is available on clinicaltrials.gov

This multicenter, open-label, 1st in human, phase 1 study (NCT03600883) evaluates safety, tolerability, PK/pharmacodynamics (PK/PD), and efficacy of AMG 510 in patients (pts) with KRASG12C advanced solid tumors. Primary endpoint: safety [eg, adverse events (AEs); dose limiting toxicities (DLT)]; key secondary endpoints: PK, ORR (overall response rate)[assessed every 6 weeks (wks)] and PFS (progression free survival). Key inclusion criteria: age ≥18; documented locally-advanced or metastatic KRASG12C; measurable or evaluable disease; ECOG ≤2; life expectancy >3 months (mo). Key exclusion criteria: active brain metastases; myocardial infarction within 6 mo. Sequential dose escalation cohorts are enrolled to evaluate safety, tolerability, PK/PD and to find the maximum tolerated dose (MTD). After identifying the MTD, 60 pts with advanced KRASG12C STs can enroll. Daily oral AMG 510 is given until disease progression (PD), intolerance, or consent withdrawal.

Clinical data presented in this manuscript was collected at participating clinical sites from September 2018 through June 2019.

Outcomes The endpoints described in this manuscript were based on RECIST 1.1 criteria for clinical responses.

## Flow Cytometry

#### Plots

Confirm that:

The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

All plots are contour plots with outliers or pseudocolor plots.

A numerical value for number of cells or percentage (with statistics) is provided.

### Methodology

Sample preparation

For in vitro studies, cells were non-enzymatically detached from the wells, washed with staining buffer (PBS/0.5% BSA), and then incubated with PE-conjugated H-2Dd, H-2Kd, or H-2Ld antibodies (BioLegend) for 30 minutes on ice. After washing, cells were resuspended in staining buffer containing SYTOX Blue Dead Cell Stain (Life Technologies), and then analyzed by flow cytometry.

For co-cultures, CellTrace Violet-labeled CD8+ T cells from spleens of OT-I transgenic mice were combined with bone marrowderived dendritic cells (BMDCs) in 96-well plates with or without further AMG 510 or MEKi treatment. Co-cultures were incubated for three days at 37°C. Cell division was assessed by flow cytometry by measuring CTV dilution in TCRβ+ CD8α+ cells.

For in vivo studies, tumors were harvested, weighed, minced, and placed in Liberase TL (0.2 mg/ml; Roche) and DNase I (20 µg/ ml; Ambion). Tumor cell suspensions were then homogenized using a gentle MACS Dissociator (Miltenyi Biotech) and incubated at 37°C for 15 minutes on a MACSmix Tube Rotator (Miltenyi Biotech). Cells were then treated with 0.02% EDTA (Sigma) and heat-inactivated FBS (ThermoFisher Scientific) and filtered to remove clumps. After centrifugation, the cell pellets were resuspended in LIVE/DEAD Fixable Blue Dead Cell Stain (ThermoFisher Scientific) for 30 minutes. Cell surface staining was performed with the indicated antibodies (see Antibodies section above) before fixation and permeabilization of the cells (Intracellular Fixation & Permeabilization Buffer Set, eBiosciences) for intracellular staining. CountBright™ Absolute Counting Beads (ThermoFisher Scientific) were added to each sample before analysis on an LSR II flow cytometer (BD Biosciences). All analyses were performed with FlowJo software v10 (FlowJo). Absolute cell counts were determined by normalizing cell numbers to beads recorded, divided by the volume of tumor aliquot analyzed and the mass of the tumor.

Instrument

In vitro samples were run on a BD LSRFortessa. In vivo samples were run on a LSR II flow cytometer (BD Biosciences).

Software

All analyses were performed with either BD FACSDiva or FlowJo software v10.

Cell population abundance

Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.

Gating strategy

For in vitro experiments, FSC-H/FSC-A gate was used to identify single cells and eliminate doublets from the analysis. FSC/SSC gate (P1) was used to gate on the population of CT-26 KRAS p.G12C cells. Cells from P1 were displayed on a histogram and SYTOX Blue negative cells were gated on to identify live cells. The mean fluorescent intensity (MFI) of MHC class I antigen expression was measured on these live cells.

For co-culture experiments, lymphocytes were first gated using FSC/SCC. Live cells were then gated using 7AAD viability dye, followed by exclusion of doublets using SSC-A/SSC-H. CD8+ T cells were then gated using fluorescently labeled antibodies. Finally, CellTrace Violet dye incorporation was assessed on the CD8+ T cells.

For in vivo experiments, cells were gated first in intact cells using FSC/SCC. Cells were then gated on live cells using the viability dye, followed by cell type-specific gating using fluorescently labeled antibodies.

| Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

# Human gut bacteria contain acquired interbacterial defence systems

https://doi.org/10.1038/s41586-019-1708-z

Received: 22 August 2018

Accepted: 20 September 2019

Published online: 30 October 2019

Benjamin D. Ross<sup>1,12</sup>, Adrian J. Verster<sup>2,12</sup>, Matthew C. Radey<sup>1</sup>, Danica T. Schmidtke<sup>1</sup>, Christopher E. Pope<sup>3</sup>, Lucas R. Hoffman<sup>1,3,4</sup>, Adeline M. Hajjar<sup>5</sup>, S. Brook Peterson<sup>1</sup>, Elhanan Borenstein<sup>2,6,7,8,9,13\*</sup> & Joseph D. Mougous<sup>1,10,11,13\*</sup>

The human gastrointestinal tract consists of a dense and diverse microbial community, the composition of which is intimately linked to health. Extrinsic factors such as diet and host immunity are insufficient to explain the constituents of this community, and direct interactions between co-resident microorganisms have been implicated as important drivers of microbiome composition. The genomes of bacteria derived from the gut microbiome contain several pathways that mediate contact-dependent interbacterial antagonism<sup>1-3</sup>. Many members of the Gramnegative order Bacteroidales encode the type VI secretion system (T6SS), which facilitates the delivery of toxic effector proteins into adjacent cells<sup>4,5</sup>. Here we report the occurrence of acquired interbacterial defence (AID) gene clusters in Bacteroidales species that reside within the human gut microbiome. These clusters encode arrays of immunity genes that protect against T6SS-mediated intra- and inter-species bacterial antagonism. Moreover, the clusters reside on mobile elements, and we show that their transfer is sufficient to confer resistance to toxins in vitro and in gnotobiotic mice. Finally, we identify and validate the protective capability of a recombinase-associated AID subtype (rAID-1) that is present broadly in Bacteroidales genomes. These rAID-1 gene clusters have a structure suggestive of active gene acquisition and include predicted immunity factors of toxins derived from diverse organisms. Our data suggest that neutralization of contact-dependent interbacterial antagonism by AID systems helps to shape human gut microbiome ecology.

Polymicrobial environments contain a plethora of biotic and abiotic threats to their inhabitants. Bacterial survival in these settings necessitates elaborate defensive mechanisms. Some of these are basal and protect against a wide range of threats, whereas others, such as CRISPR-Cas, represent adaptations that are unique to the specific threats encountered by a bacterial lineage<sup>6,7</sup>. In densely colonized habitats such as the mammalian gut, overcoming contact-dependent interbacterial antagonism is a major hurdle to survival. The T6SS is a pathway widely used by gut bacteria to mediate the delivery of toxic effector proteins into neighbouring cells<sup>8-10</sup>. Although kin cells are innately resistant to these effectors via cognate immunity proteins, whether non-self cells in the gut can escape intoxication is unknown. Notably, several recent studies have reported that bacteria from a range of environments can encode T6S immunity genes that lack an accompanying cognate effector gene<sup>2,9,11-13</sup>. It has been suggested that these genes are involved in  $interbacterial\, defence, but they \, have \, yet \, to \, be \, functionally \, investigated$ in a physiological context.

## B. fragilis T6S immunity in the human gut microbiome

To identify potential mechanisms of defence against T6S-delivered interbacterial effectors in the human gut, we mined a large collection of shotgun metagenomic samples (n = 553) from several studies of the human gut microbiome for sequences homologous to known immunity genes<sup>14,15</sup> (Supplementary Table 1). We first focused our efforts on *Bacteroides fragilis*, which has a well-described and diverse repertoire of effector and cognate immunity genes<sup>3,8,9</sup> (Supplementary Table 2). As expected for genes predicted to reside within the *B. fragilis* genome, sequences mapping to these immunity loci were detected at a similar abundance to that of *B. fragilis* species-specific marker genes in many microbiome samples (Fig. 1a, grey; Supplementary Table 1). However, in a second subset of samples, immunity genes were detected at a significantly higher (more than ten times) abundance than expected given the abundance of *B. fragilis* (Fig. 1a, blue). Finally, we identified a third subset of samples in which the

Department of Microbiology, University of Washington, Seattle, WA, USA. Department of Genome Sciences, University of Washington, Seattle, WA, USA. Department of Genome Sciences, University of Washington, Seattle, WA, USA. Department of Computer Science and Engineering, University of Washington, Seattle, WA, USA. Beattle, WA, USA. Beat

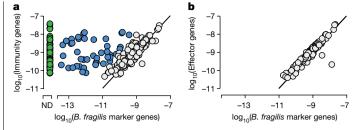


Fig. 1 | T6SS orphan immunity genes are found in human gut microbiomes. a, b, Comparison of abundance of B. fragilis-specific T6SS immunity genes (a) or effector genes (b) with B. fragilis marker genes in adult microbiome samples derived from the Human Microbiome Project (HMP) and METAgenomics of the Human Intestinal Tract (MetaHIT) studies (Supplementary Table 1). Abundance values denote the number of reads mapped to the gene, normalized by gene length and total number of reads in the sample. Abundances are calculated as the average abundance of all immunity, effector or B. fragilis species-specific marker genes. Samples with undetectable B. fragilis (green) and samples in which immunity gene abundance exceeds that of Bacteroides by over tenfold (blue) are highlighted. ND, not detected.

sequences of immunity genes were detected in the absence of B. fragilis (Fig. 1a, green). These latter sequences include close homologues of 12 out of 14 unique immunity genes (i1-i14) encoded by B. fragilis (Extended Data Fig. 1a). In contrast to the pattern observed for immunity genes, we found no samples in which the abundance of corresponding cognate effector genes markedly exceeded that of B. fragilis (Fig. 1b, Supplementary Table 1).

#### Orphan immunity genes are encoded by many species

The detection of *B. fragilis* immunity gene homologues in samples in which we were unable to detect B. fragilis strongly suggests that these elements are encoded by other bacteria in the gut. Indeed, BLAST searches revealed the genomes of several Bacteroides spp. that contain B. fragilis T6S immunity gene homologues, including B. ovatus (i6, i7, i5 and i14), B. vulgatus (i8 and i13), B. helcogenes (i1, i9 and i10) and B. coprocola (i8 and i13). To determine whether these bacteria could also account for the presence of immunity genes in the human gut microbiome, we assembled full-length predicted immunity genes from the metagenomic sequencing reads of individual microbiomes. We limited this assembly to homologues of i6—the most prevalent immunity gene detected in samples lacking B. fragilis (Extended Data Fig. 1a). Clustering of the recovered homologues showed that most i6 sequences distribute into three discrete clades that differ by several nucleotide substitutions (i6:cl, cll and clll) (Fig. 2a and Supplementary Table 3). A comparison of these immunity sequences to available bacterial genomes revealed a clade matching cognate immunity genes in B. fragilis (i6:cl). In addition, we found an i6 sequence homologous to i6:cIII in the genome of B. ovatus, which we previously found does not contain the cognate T6SS effector gene8.

To identify the species that encode these sequences in human gut microbiomes, we used a simplified linear model to identify Bacteroides spp., the abundance of which in microbiome samples best fits that of each immunity sequence clade. We found that i6:cIII is best explained in gut metagenomes by B. ovatus (Fig. 2b, c), which suggests that although reference genomes of both B. fragilis and B. ovatus contain these sequences, it is most often contained by the latter in natural populations. We could not confidently define a single species containing i6:cII by this method (Fig. 2d, e); therefore, we applied the same analysis pipeline to an infant microbiome dataset for which matching stool samples were available (Extended Data Fig. 1b and Supplementary Table 4). Whole-genome sequencing of isolates identified B. xylanisolvens as a bacterium containing i6:cII in these samples. Notably, this species best fit the abundance of i6:cII genes in the large adult metagenomic

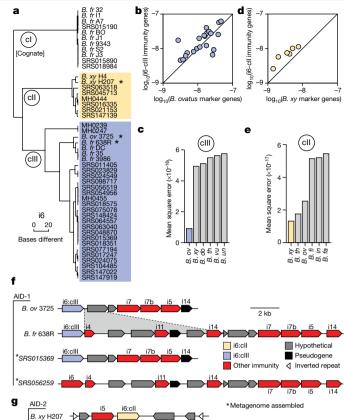
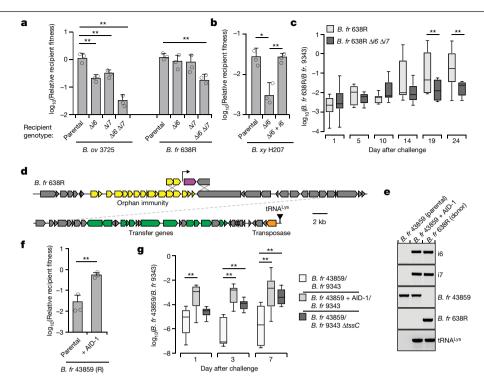


Fig. 2 | T6SS orphan immunity gene clusters are encoded by several species in the human gut microbiome. a, Dendrogram depicting hierarchical clustering of orphan immunity gene i6 sequences extracted from genomes (n=15) and metagenomes (n=32) derived from the indicated HMP (SRS) or MetaHIT (MH) samples. Sequence clades are denoted cI-III. Asterisks indicate strains shown in f and g, b-d. Comparison of abundance of genes from the indicated immunity clades (colours as in a) with marker genes from B. ovatus (b) or B. xylanisolvens (d) in adult microbiome samples (Supplementary Table 1). c, e, Linear model error values for the six species best fitting i6:cIII (c) and i6:cII (e) gene abundances calculated as in Fig. 1.  $\mathbf{f}$ ,  $\mathbf{g}$ , Representative AID-1( $\mathbf{f}$ ) and AID-2 (g) gene clusters containing homologues of the indicated B. fragilis T6S immunity genes. The i6 gene of SRS056259 did not meet our sequence depth coverage requirements for inclusion in i6:cIII. The B. xylanisolvens strain in g was sequenced as a part of this study (BioProject PRJNA484981). A region of difference between B. ovatus 3725 and B. fragilis 638R clusters is highlighted. All strain abbreviations are defined in Supplementary Table 3.

datasets, albeit by a narrow margin (Fig. 2d). On the basis of these observations, we hypothesized that orphan immunity genes—encoded by B. fragilis and other species of Bacteroides—have an adaptive role in the gut by providing defence during intra- and inter-species antagonism.

## AID system immunity genes neutralize T6S toxins

To gain insight into the function of orphan immunity genes, we examined their genomic context in available reference genomes and assembled sequence scaffolds from metagenomic data. We found that homologues of B. fragilis T6S immunity genes i6 and i7 are located together within discrete gene clusters, which we termed AID-1 (acquired interbacterial defence 1) systems, in several Bacteroides strains and in microbiome samples (Fig. 2f). Within the AID-1 gene cluster, we identified distant homologues and pseudogenized remnants of additional B. fragilis immunity genes, including i4, i5, i11 and i14 (Fig. 2f and Supplementary Table 5). These findings prompted us to search for more distant homologues of B. fragilis T6S immunity genes. This revealed



**Fig. 3** | **Orphan immunity genes are mobile and protect against T6S-delivered toxins. a, b,** Outcomes of growth competition assays between the indicated strains containing AID-1 (a) or AID-2 (b) versus B. fragilis 9343. Relative recipient fitness was determined by calculating the ratio of final to initial colony-forming units (c.f.u.) and normalizing to corresponding experiments with T6S-inactive B. fragilis 9343 ( $\Delta tssC$ ). Data are mean  $\pm$  s.d. of three technical replicates indicative of at least three biological replicates.  $^*P < 0.05$ ,  $^*P < 0.01$ , unpaired two-tailed t-test.  $\mathbf{c}$ , Outcome of pairwise competition between the indicated

 $B. fragilis\, strains\, in\, germ-free\, mice.\, Mice\, were\, colonized\, with\, B.\, fragilis\, 9343\, for\, one\, week\, and\, challenged\, with\, 638R\, (n=8\, mice\, per\, group\, for\, each\, of\, two$ 

independent experiments). For box plots, the middle line denotes the mean for each group; the box denotes the interquartile range; and the whiskers denote the minimum and maximum values. \*\*P<0.01, Mann–Whitney U-test for each time point.  $\mathbf{d}$ , Schematic depicting a B-fragilis 638R ICE containing the AID-1 cluster depicted in Fig. 2f.  $\mathbf{e}$ , PCR analysis of an AID-1 transfer experiment. Schematic with primer locations provided in Extended Data Fig. 3f. Transfer data are representative of two independent biological replicates.  $\mathbf{f}$ , Outcomes of growth competition assays between B-fragilis 43859 or a derivative AID-1ICE recipient and B-fragilis 9343. Data presentation and statistics are as in  $\mathbf{a}$  and  $\mathbf{b}$ .  $\mathbf{g}$ , Results of pairwise competition between the indicated B-fragilis strains in germ-free mice (n = 6 mice from two independent experiments). Statistics and

gene clusters containing orphan homologues of i1, i3, i8, i9, i10 and i13 in diverse *Bacteroides* genomes and we also identified distant homologues of these genes in a metagenomic gene catalogue (Extended Data Fig. 2a, b). In *B. xylanisolvens*, we found that genes belonging to i6:cll are located in a unique, but analogous context adjacent to a homologue of i5 on an apparent transposable element<sup>17</sup> (Fig. 2g). We designated this sequence AID-2.

We next defined the phenotypic implications of orphan immunity genes of Bacteroides spp. during interbacterial competition. B. fragilis 9343 encodes the cognate effectors for i6 and i7, and previous data demonstrate that the corresponding toxins antagonize assorted Bacteroides spp. in vitro and in gnotobiotic mice<sup>9</sup>. We thus used this strain in growth competition assays against Bacteroides spp. bearing orphan immunity genes, derivative strains containing deletions of these genes, or genetically complemented strains. These experiments showed that in both B. ovatus and B. fragilis, AID-1 system genes grant immunity against corresponding T6S effectors (Fig. 3a, Extended Data Fig. 3a, b). The i6 and i7 genes of *B. ovatus* did not influence the outcome of its competition with B. fragilis 638R, which possesses an orthogonal effector repertoire (Extended Data Fig. 3c). Finally, we also found that an i6:cll gene from a B. xylanisolvens AID-2 system gives this bacterium protection against e6 of B. fragilis 9343 (Fig. 3b). Together, these data show that the orphan immunity genes of several Bacteroides species—localized to AID systems—can confer protection against effectors delivered by the T6SS of B. fragilis.

*B. fragilis* is typically found as a clonal population in the human gut microbiome, and recent studies suggest that this is in part owing to

active strain exclusion via the T6SS $^{8.18}$ . However, in colonization experiments in gnotobiotic mice, certain B.fragilis strain pairs inexplicably coexist $^{10}$ . We noted that one such pair corresponds to B.fragilis 9343 and B.fragilis 638R, the latter of which contains an AID-1 system containing homologues of i6 and i7. To determine whether our in vitro results with these strains extend to a more physiological setting, we measured the fitness contribution of the orphan immunity genes encoded by B.fragilis 638R after pre-colonization of germ-free mice with B.fragilis 9343 (Fig. 3c, Extended Data Fig. 3d, e). Our results indicated that the cumulative protection afforded by orphan i6 and i7 genes underlies the ability of B.fragilis 638R to persist during T6S-mediated antagonism in vivo.

#### AID-1 transfer confers protection against T6S

Notably, we found that AID-1 resides on a predicted mobile integrative and conjugative element (ICE), which provides a possible explanation for its distribution <sup>19</sup> (Fig. 3d). To test whether this element can be transferred between strains, we performed mobilization studies using *B. fragilis* 638R as a donor and *B. fragilis* 43859 as a recipient. An antibiotic-resistance marker was inserted within AID-1 to facilitate the detection of its transfer. With this tool, we readily detected AID-1 transfer (Fig. 3e). This occurred at a frequency of approximately  $5 \times 10^{-6}$ , in line with previous quantification of ICE mobility in *Bacteroides* spp. <sup>20</sup>. Next, we asked whether the transfer of AID-1 to *B. fragilis* 43859 is sufficient to confer resistance to T6S-mediated antagonism. In vitro growth competition assays against *B. fragilis* 9343 showed that AID-1 effectively neutralizes intoxication by e6 and e7 (Fig. 3f). The receipt of AID-1 also

granted notable protection to *B. fragilis* 43859 against T6S-mediated killing in germ-free mice pre-colonized with *B. fragilis* 9343 (Fig. 3g). Together, these findings indicate that the transfer of a mobile orphan immunity island to a naive *Bacteroides* strain is sufficient to provide defence against T6S effectors.

Deciphering the contribution of individual gene products, or even whole pathways, to bacterial fitness in complex microbial communities is challenging. We reasoned that the identification of orphan immunity genes in human gut metagenomes, coupled with our ability to infer their organismal source, provided an opportunity to measure the effect of these defensive factors on competitiveness in the gut. To this end. we compared the abundance of B. ovatus strains with and without i6 and i7 orphan immunity genes in human gut metagenome samples. We found that the average abundance of B. ovatus strains with orphan immunity genes significantly exceeds that of those without orphan genes (Extended Data Fig. 3g). One interpretation of this finding is that the acquisition of i6 and i7 allows B. ovatus to increase its niche: however. there are several potential caveats inherent to these correlative data that cannot be ruled out. For example, B. ovatus strains that contain i6 and i7 might be related and enriched for other fitness determinants that account for their abundance.

#### Recombinase-associated AID systems are prevalent

Given the benefit of orphan immunity genes against B. fragilis effectors, we hypothesized that this mechanism of inhibiting interbacterial antagonism should extend to effectors produced by other species. We previously found that B. fragilis is antagonized by other Bacteroides spp. in the human gut microbiome<sup>8</sup>. In addition to the T6SS present exclusively in B. fragilis, this species and other Bacteroides species can possess other T6SSs, referred to as GA1 (genetic architecture 1) and GA2, with a distinct and non-overlapping repertoire of effector and immunity genes<sup>3</sup>. The effectors of these systems exhibit hallmarks of antibacterial toxins and we demonstrated their ability to mediate interbacterial antagonism<sup>2,8</sup> (Fig. 4, Extended Data Fig. 4). Therefore, we searched *B. fragilis* genomes for sequences that are homologous to the immunity genes corresponding to these systems. In 29 out of the 122 available B. fragilis genomes, we identified apparent orphan homologues of these immunity genes grouped within gene clusters (Fig. 4a). Although analogous to the AID-1 and AID-2 systems, these clusters have several unique characteristics including skewed GC content, conservation of a gene encoding a predicted XerD-family tyrosine recombinase, and repetitive intergenic sequences<sup>21</sup> (Extended Data Fig. 5a-c). These often-large gene clusters, hereafter referred to as recombinase-associated AID (rAID-1) systems. can exceed 16 kb and contain up to 31 genes with varying degrees of homology to T6S immunity genes and predicted immunity genes associated with other interbacterial antagonism pathways<sup>2</sup> (Fig. 4a, Extended Data Fig. 5d, e, Supplementary Table 6). Using the shared characteristics of B. fragilis rAID-1 systems, we searched for related gene clusters across sequenced Bacteroidales genomes. We found that more than half of sequenced bacteria belonging to this order possess a rAID-1 system (226 out of 423) (Fig. 4b, Supplementary Tables 6, 7). In summary, these gene clusters contain 579 unique genes, which encompass homologues of 25 Bacteroidales T6S immunity genes.

The prevalence of rAID-1 genes in Bacteroidales genomes suggested that these elements may be common in the human gut microbiome. To investigate this, we searched metagenomic data for sequences that map to Bacteroidales T6S orphan immunity genes found within rAID-1 systems. Notably, we found one or more rAID-1 immunity genes in 551 out of 553 samples using a 97% sequence identity threshold to map reads (Supplementary Table 8). These rAID-1 immunity genes diverge considerably from corresponding cognate immunity genes (corresponding to 32–91% amino acid identity), which suggests that the latter are an unlikely source of significant false positives in this analysis. We also searched the same samples for rAID-1-associated recombinase

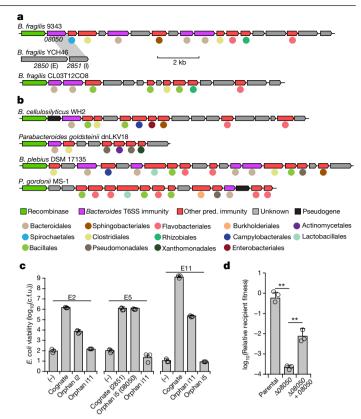


Fig. 4 | rAID systems encode toxin-neutralizing immunity genes and are prevalent in human gut microbiomes. a, b, rAID-1 clusters from the indicated B. fragilis (a) or Bacteroidales (b) species. rAID-1 genes were assigned to functional immunity classes (indicated by gene colouring) via profile HMM scans and BLAST against a curated database of Bacteroidales T6SS immunity genes<sup>2,8</sup>. Coloured circles indicate taxonomic association of the top non-rAID-1 homologue. Homology (70% amino acid identity) between gene 08050 of the  $\textit{B.fragilis}\,9343\,\text{rAID-1}\,\text{cluster}\,\text{and}\,\text{a}\,\text{T6S}\,\text{cognate}\,\text{immunity}\,\text{gene}\,\text{from}\,\textit{B.fragilis}$ YCH46 (2851) is indicated. c, Viable E. coli cells recovered from cultures expressing the indicated proteins (see Supplementary Table 10 for locus tags). Data are mean  $\pm$  s.d. of three technical replicates representative of three independent biological replicates. d, Outcomes of growth competition assays between the indicated *Bacteroides* strains (n = 3 biologically independent)samples). The relevant rAID-1 gene of B. fragilis 9343 and its corresponding effector within B. fragilis YCH46 are depicted in a. \*\*P < 0.01, unpaired two-tailed t-test. Data are mean  $\pm$  s.d.

sequences. Although recombinase genes are widely distributed across bacteria, close homologues (more than 50% amino acid identity) of those found associated with rAID-1 systems are restricted to this context and only found in Bacteroidales genomes. Consistent with this, we found that the abundance of rAID-1 recombinase genes correlates strongly with the genus Bacteroides (Extended Data Fig. 5f).

#### Divergent rAID-1 immunity genes protect against T6S

Orphan immunity genes encoded within rAID-1 clusters diverge more from cognate immunity than do those within AID-systems. Thus, we sought to experimentally validate the ability of rAID-1 immunity genes to protect bacteria from intoxication. Because most bacteria containing rAID-1 systems have limited genetic tools, we used *Escherichia coli* to identify three Bacteroidales T6S effector genes that intoxicate cells in a manner that is neutralized by cognate immunity. In each case, we found that co-expression of these effector genes with corresponding rAID-1-associated orphan immunity genes, but not mismatched orphan immunity genes, restored *E. coli* growth (Fig. 4c). Both of the genes from

one effector-orphanimmunity pair that we validated derive from genetically tractable strains; B. fragilis YCH46 (effector, 2850) and B. fragilis 9343 (orphan immunity, 08050). In vitro growth competition experiments with these strains, and mutant and genetically complemented derivatives, showed that an endogenous rAID-1 orphan immunity gene of B. fragilis 9343 can neutralize a T6S-delivered toxin (Fig. 4d).

The orphan immunity systems that we defined consist of many genes and their expression could incur a substantial metabolic burden. As a first step towards understanding the regulation of AID systems, we performed quantitative PCR with reverse transcription (qRT-PCR) analysis to compare the expression of the systems in the presence and absence of a competitor strain. These studies provided evidence that transcription of both systems is induced by co-cultivation with a competitor strain (Extended Data Fig. 5g). We also examined meta-transcriptomic data for evidence of AID expression. Owing to a paucity of such data available for samples definitively containing AID-1 and AID-2, we could not systematically quantify the expression of these systems. However, using conservative criteria for defining rAID-1-associated genes in metatranscriptomic data, we found evidence for the expression of this system in every sample derived from a large study  $(n=156)^{22}$  (Supplementary Table 9). In some samples, such as those with high levels of Bacteroides, rAID-1 genes accounted for nearly 1 in 10,000 of all meta-transcriptomic reads. Together with our functional characterization of AID systems, these findings suggest that acquisition and maintenance of consolidated orphan immunity determinants is a common mechanism by which Bacteroidales defend against interbacterial antagonism in the human gut microbiome.

#### Discussion

Mounting evidence suggests that competitive interactions between bacteria predominate in many environments<sup>23</sup>. This evolutionary pressure has undoubtedly led to the wide dissemination of idiosyncratically orphaned immunity genes that are predicted to provide resistance to diverse antagonistic pathways<sup>2,11,24,25</sup>. Modelling studies predict that interbacterial antagonism is a crucial contributor to the maintenance of a stable gut community26. Our findings reveal that a corollary of the pervasiveness of antagonistic mechanisms is strong selective pressure for genes that can provide protection against attack, establishing a molecular arms race that has led to the diversification and expansion of T6S effectors. Deciphering the link between orphan immunity genes and the bacteria harbouring the cognate effectors may help to shed light on the physical connectivity of bacteria in the gut microbiome.

It is now appreciated that phage defence mechanisms, including the adaptive system CRISPR, are crucial for bacteria to cope with the omnipresent threat and deleterious outcome of phage infection<sup>7</sup>. However, the ubiquity of interbacterial antagonistic systems suggests that in most habitats, bacteria are equally, or perhaps more likely to be subject to attack and potential cell death via the action of other bacteria<sup>2</sup>. Our characterization of AID systems encoded by prevalent members of the human gut microbiota seems to reconcile these observations and demonstrate that the neutralization of contact-dependent interbacterial antagonism can be a critical mechanism for survival in polymicrobial environments. In addition, it suggests that analogous to the immune system of vertebrates, that of bacteria includes arms specialized in viral or bacterial defence.

#### Online content

Any methods, additional references. Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1708-z.

- Whitney, J. C. et al. A broadly distributed toxin family mediates contact-dependent antagonism between Gram-positive bacteria. eLife 6, e26938 (2017).
- Zhang, D., de Souza, R. F., Anantharaman, V., Iyer, L. M. & Aravind, L. Polymorphic toxin systems: comprehensive characterization of trafficking modes, processing, mechanisms of action, immunity and ecology using comparative genomics. Biol. Direct 7, 18 (2012).
- Coyne, M. J., Roelofs, K. G. & Comstock, L. E. Type VI secretion systems of human gut Bacteroidales segregate into three genetic architectures, two of which are contained on mobile genetic elements. BMC Genomics 17, 58 (2016).
- Russell, A. B. et al. A type VI secretion-related pathway in Bacteroidetes mediates interbacterial antagonism. Cell Host Microbe 16, 227-236 (2014).
- Hood, R. D. et al. A type VI secretion system of Pseudomonas aeruginosa targets a toxin to bacteria, Cell Host Microbe 7, 25-37 (2010).
- Cornforth, D. M. & Foster, K. R. Competition sensing: the social side of bacterial stress responses, Nat. Rev. Microbiol. 11, 285-293 (2013).
- Hille, F. et al. The biology of CRISPR-Cas: backward and forward. Cell 172. 1239-1259 (2018)
- Verster, A.J. et al. The landscape of type VI secretion across human gut microbiomes reveals its role in community composition. Cell Host Microbe 22, 411-419 (2017).
- Wexler, A. G. et al. Human symbionts inject and neutralize antibacterial toxins to persist in the gut. Proc. Natl Acad. Sci. USA 113, 3639-3644 (2016).
- Hecht, A. L. et al. Strain competition restricts colonization of an enteric pathogen and prevents colitis. EMBO Rep. 17, 1281-1291 (2016).
- 11. Kirchberger, P. C., Unterweger, D., Provenzano, D., Pukatzki, S. & Boucher, Y. Sequential displacement of type VI secretion system effector genes leads to evolution of diverse immunity gene arrays in Vibrio cholerae. Sci. Rep. 7, 45133 (2017).
- Steele, M. I., Kwong, W. K., Whiteley, M. & Moran, N. A. Diversification of type VI secretion system toxins reveals ancient antagonism among bee gut microbes. mBio 8, e26938 (2017).
- Ting, S. Y. et al. Bifunctional immunity proteins protect bacteria against Ftsz-targeting ADP-ribosylating toxins. Cell 175, 1380-1392 (2018).
- Lloyd-Price, J. et al. Strains, functions and dynamics in the expanded Human Microbiome Project. Nature 550, 61-66 (2017).
- Qin, J. et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464, 59-65 (2010).
- 16. Manor, O. et al. Metagenomic evidence for taxonomic dysbiosis and functional imbalance in the gastrointestinal tracts of children with cystic fibrosis, Sci. Rep. 6, 22493 (2016).
- Siguier, P., Gourbeyre, E. & Chandler, M. Bacterial insertion sequences: their genomic impact and diversity, FEMS Microbiol, Rev. 38, 865-891 (2014).
- Zhao, S. et al. Adaptive evolution within gut microbiomes of healthy people. Cell Host Microbe 25, 656-667, e8 (2019).
- Wozniak, R. A. & Waldor, M. K. Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. Nat. Rev. Microbiol. 8, 552-563 (2010)
- Stevens, A. M., Shoemaker, N. B. & Salyers, A. A. The region of a Bacteroides conjugal chromosomal tetracycline resistance element which is responsible for production of plasmidlike forms from unlinked chromosomal DNA might also be involved in transfer of the element, J. Bacteriol, 172, 4271-4279 (1990).
- Castillo, F., Benmohamed, A. & Szatmari, G. Xer site specific recombination: double and single recombinase systems. Front. Microbiol. 8, 453 (2017)
- Abu-Ali, G. S. et al. Metatranscriptome of human faecal microbial communities in a cohort of adult men. Nat. Microbiol. 3, 356-366 (2018).
- Foster, K. R. & Bell, T. Competition, not cooperation, dominates interactions among culturable microbial species. Curr. Biol. 22, 1845-1850 (2012).
- Poole, S. J. et al. Identification of functional toxin/immunity genes linked to contactdependent growth inhibition (CDI) and rearrangement hotspot (Rhs) systems. PLoS Genet. 7. e1002217 (2011).
- Drider, D., Fimland, G., Héchard, Y., McMullen, L. M. & Prévost, H. The continuing story of class IIa bacteriocins, Microbiol, Mol. Biol, Rev. 70, 564-582 (2006).
- 26. Coyte, K. Z., Schluter, J. & Foster, K. R. The ecology of the microbiome: networks, competition, and stability, Science 350, 663-666 (2015)

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

#### **Methods**

#### Microbiome data

Metagenomic data from healthy adults were obtained from several large-scale sequencing projects. We specifically obtained 147 samples from the Human Microbiome Project (HMP) 1.0, 100 samples from HMP 1.2, and 99 and 207 samples from two different MetaHIT datasets  $^{14.15,27.28}$ . We further obtained paired metagenomic–metatranscriptomic data from a study of 156 individuals  $^{22}$ . Finally, we obtained a database of genes identified from 1,267 assembled metagenomes as part of the integrated gene catalogue (IGC)  $^{29}$ .

#### Analysis of gene and species abundances in microbiome samples

We previously compiled a list of T6SS immunity and effector genes<sup>8</sup>. We also compiled a list of species-specific marker genes for all Bacteroides species obtained from MetaPhlAn 2.030. To determine the abundance of a given immunity, effector or marker gene in each metagenomic sample, single-end metagenomic reads were aligned to gene sequences using bowtie2, allowing for one mismatch in the seed31. We counted the number of reads that aligned to each such gene with at least 80% nucleotide identity (to encompass divergent orphan immunity gene sequences) and minimum mapping quality of 20. The abundance of a gene was calculated as the number of reads aligned to this gene, normalized by the gene length and by the library size. For each species, the average gene level abundance of all species-specific marker genes was used to assess the species abundance. For the total Bacteroides abundance, we used the sum of all species-specific marker genes in the genus. Samples were only included in an analysis if they had at least 10 reads mapping to the T6SS genes in question (effectors, immunity or recombinases). On the basis of the abundance of GA3 immunity genes and B. fragilis, we split samples into those in which B. fragilis was not detectable, those in which the immunity gene had more than ten times the abundance of the B. fragilis marker gene, and those in which such discrepancy between the abundance of immunity genes and that of B. fragilis was not observed. Meta-transcriptomics data were processed similarly to metagenomics data, except that abundance values were converted to a reads per kilobase of transcript per million mapped reads (RPKM) value for familiarity with canonical RNA sequencing analysis.

#### Orphan immunity phylogenetic analysis

Filtered reads derived from human shotgun microbiome datasets were aligned using bowtie2 as described above and subsequently converted to a pileup using samtools with parameters --excl-flags UNMAP,QCFAIL,DUP -A -q0 -C0 -B<sup>31,32</sup>. A sequence corresponding to the most abundant version of the immunity gene in the sample was reconstructed from that pileup as follows. First, 50 bases from the start and end were trimmed due to a propensity for low coverage. Second, at all sites with at least 10 times coverage the base was set to the major allele. Sites with less than ten times coverage were assigned an ambiguous base. Finally, we only kept the reconstructed sequence in metagenomic samples where at least 90% of the sequence had more than ten times coverage. The number of single nucleotide polymorphisms between all immunity sequences, both from metagenomic samples and from Bacteroides genomes, was calculated and used to populate a distance matrix. Because obtained distances were small (for example, a single base difference), we used hierarchical clustering (with complete linkage), rather than standard phylogenetic reconstruction methods, to visualize the relatedness between different sequences. Sequence clades defined by hierarchical clustering are denoted (cI-III), as discussed in the text.

#### Assigning orphan immunity sequences to bacterial species

We aimed to identify the species most likely to encode the immunity gene in each cluster of identical sequences reconstructed from metagenomes. Only clusters with at least three sequences were used, to ensure statistical confidence. The abundance of each species was assessed based on

species-specific marker genes as described above. We specifically used a simple linear model that assumed that only a single species encodes the immunity gene. We further assumed a one-to-one relationship between species marker gene abundance and orphan immunity gene abundance, and accordingly fixed the intercept at zero and allowed a single species with a slope of one. The fit of the model for each species was calculated as the mean square error over all samples. The most likely species to encode the immunity gene was determined by the minimum mean squared error.

#### Assembly of orphan immunity sequences from metagenomes

Paired-end metagenomic sequencing data were assembled using Soap-DeNovo2 with a kmer length of 63 and an average insert size of 200<sup>33</sup>. BLAST was used to identify the contig that contained the orphan immunity gene, and GeneMarkS was used to predict protein-coding genes<sup>34</sup>.

#### **Bacterial culture conditions**

Anaerobic culturing procedures were performed either in an anaerobic chamber (Coy Laboratory Products) filled with 70%  $N_2$ , 20%  $CO_2$  and  $10\%\,H_2$ , or in Becton Dickson BBL EZ GasPak chambers. *E. coli* EC100D  $\lambda$  pir and S17-1 $\lambda$  pir strains were grown aerobically at 37 °C on lysogeny broth (LB) agar. Unless otherwise noted, Bacteroides strains were cultured under anaerobic conditions on brain heart infusion (BHI) agar (Becton Dickinson) supplemented with 50  $\mu$ g ml $^{-1}$  haemin and 10% sodium bicarbonate (BHIS) $^{35}$ . Antibiotics and chemicals were added to media as needed at the following concentrations: trimethoprim 50  $\mu$ g ml $^{-1}$ , carbenicillin 150  $\mu$ g ml $^{-1}$ , gentamicin 15  $\mu$ g ml $^{-1}$  (*E. coli*), gentamicin 60  $\mu$ g ml $^{-1}$  (*Bacteroides*), erythromycin 12.5  $\mu$ g ml $^{-1}$ , tetracycline 6  $\mu$ g ml $^{-1}$ , chloramphenicol 12  $\mu$ g ml $^{-1}$ , floxuridine (FUdR) 200  $\mu$ g ml $^{-1}$ .

#### **Genetic techniques**

Standard molecular procedures were used for the creation, maintenance and E. coli transformation of plasmids. All primers used in this study were  $synthesized \, by \, Integrated \, DNA \, Technologies \, (IDT). \, Phusion \, polymerase, \, and \, continuous \, and \, continuous \, depends on the property of the p$ restriction enzymes, T4 DNA ligase, and Gibson Assembly Reagent were obtained from New England Biolabs (NEB). A comprehensive list of primers, plasmids, and strains are provided (Supplementary Table 10). Deletion of the gene encoding thymidine kinase in B. fragilis, B. ovatus and B. xylanisolvens strains was performed by cloning respective genomic flanking regions into the vector pKNOCK as previously described<sup>36</sup>. In brief, pKNOCK-tdk plasmids were mobilized into Bacteroides strains via overnight aerobic mating with E. coli. Integrants were isolated by plating on selective media, were passaged once without antibiotics to allow for plasmid recombination, and plated for counter selection on FUdR. Recovered single colonies were patched onto selective media to ensure loss of pKNOCK, and disruption of tdk was confirmed by PCR. Subsequent deletion of orphan immunity genes was performed in  $\Delta t dk$ strains via a similar counter selection strategy, except employing the suicide plasmid pExchange in place of pKNOCK<sup>9</sup>. Genomic deletions were confirmed by PCR. Gene complementation was performed by cloning genes into pNBU2-erm us1311 for constitutive expression<sup>37</sup>.

#### Isolation of Bacteroides strains from faecal samples

Faecal samples from healthy infants used for strain isolation were collected as part of a previous study approved by the Seattle Children's Hospital Institutional Review Board  $^{16,38}$ . Frozen stool samples stored at  $-80\,^{\circ}\text{C}$  were manually homogenized, serially diluted in tryptone yeast glucose (TYG) broth, and plated under anaerobic conditions on Bacteroides bile esculin (BBE) agar plates (Oxyrase). Single colonies that exhibited esculin hydrolysis as indicated by the production of black pigment on BBE agar were sub-cultured in TYG broth with the addition of 60  $\mu g \, \text{m} \, \text{l}^{-1}$  gentamic in until stationary phase and then were frozen at  $-80\,^{\circ}\text{C}$  after the addition of sterile glycerol to 20% final concentration. Single colonies isolated from these stocks were subsequently screened by PCR with primers targeting the orphan i6 gene as assembled from metagenomic short read sequence data  $^{16}$ .

#### **Genome sequencing**

Genomic DNA used for Illumina sequencing was prepared by collecting *Bacteroides* strains grown overnight on BHIS blood agar plates. Cells resuspended from plates were washed in PBS before DNA extraction with the Qiagen DNeasy Blood and Tissue Kit. Sequencing was performed on an Illumina MiSeq using the V3 Reagent kit at the Northwest Genomics Center sequencing facility at the University of Washington. AID clusters often appear in highly repetitive genomic contexts (for example, mobile elements) and are often split into multiple scaffolds in reference genomes. To compensate for this, we also performed long-read sequencing via PacBio on a subset of genomes. To this end, high molecular mass DNA was extracted using the Qiagen Genomic-tip Kit and sequenced by SNPsaurus using a PacBio Sequel. Hybrid long read and short read assemblies were conducted using Unicycler<sup>39</sup>. Species identification was performed by blast searches with species-specific marker genes<sup>30</sup>.

#### Interbacterial competition assays

Bacteroidales strains were grown on BHIS blood agar plates overnight at 37 °C. Bacteria were resuspended from plates in BHIS broth and the optical density (OD) of each strain was adjusted to a 10:1 B. fragilis NCTC 9343 to competitor ratio ( $OD_{600}$  6.0 to 0.6) for competitions involving B. xylanisolvens and B. ovatus, or 1:1 ratio for competitions involving B. fragilis 638R ( $OD_{600}$  6.0). Equal volumes of each strain at the adjusted OD were mixed and 5 µl of bacterial mixtures were spotted onto predried BHIS blood agar plates, in triplicate spots. Competitions were allowed to proceed for 20-24 h at 37 °C under anaerobic conditions before spots were collected into BHIS broth. Competition outcomes were quantified in one of two ways: (1) by serial dilution for enumeration of c.f.u. after plating on BHIS-selective plates containing either erythromycin or tetracycline; or (2) purification of total genomic DNA using the Qiagen DNeasy Blood and Tissue Kit and subsequent quantification by qPCR using strain-specific primers (see Supplementary Table 10). For antibiotic selection, B. fragilis 9343 was marked with erythromycin resistance by integration of pNBU2-erm at the att1 site<sup>37</sup>. Other strains were either naturally tetracycline resistant, or marked by integration of pNBU2-tet-BCO1. Strains with insertions of pNBU2 were selected for matching integration sites by PCR with primers flanking att loci<sup>40</sup>. Interbacterial competitions between strains of *B. fragilis* occasionally exhibited T6SS-independent phenotypes that were dependent on the initial starting ratio of the strains used<sup>41</sup>.

#### Interbacterial mobile element transfer assays

Allelic exchange was used to engineer a high-expression chloramphenicol resistance cassette onto the AID-1 system of B. fragilis 638R, replacing BF638R\_2056-2058 $^{42}$ . Chloramphenicol-resistant B. fragilis 638R cells were mixed on BHIS blood agar plates with erythromycin-resistant B. fragilis ATCC 43859 cells at a 1:1 ratio (OD $_{600}$  6.0). After overnight coculture, bacterial mixtures were collected and plated on BHIS plates containing either erythromycin alone (to quantify c.f.u. of total ATCC cells), or erythromycin and chloramphenicol (to quantify c.f.u. of AID-1 recipient ATCC cells). Double-resistant colonies were screened individually by PCR to confirm strain identity, the presence of the AID-1 system, and the genomic integration site at a tRNA<sup>Lys</sup> gene (see Supplementary Table 10 for primers used).

#### **Gnotobiotic mice studies**

Germ-free 6–12-week-old female Swiss Webster mice from several litters were randomized, housed simultaneously in pairs in single Techniplast cages with a 12-h light/dark cycle, and fed a standard laboratory diet (Laboratory Autoclavable Rodent Diet 5010, LabDiet), in accordance with guidelines approved by the University of Washington Institutional Animal Care and Use Committee. Blinding was not performed, and no statistical methods were used to determine sample size. Reasonable

numbers of animals were used considering limitations of housing and maintenance under gnotobiotic conditions. Bacteroides fragilis strains were introduced into mice via oral gavage of 10<sup>8</sup> c.f.u. suspended in 0.2 ml of sterile PBS with 20% glycerol. Challenge with B. fragilis 638R or B. fragilis ATCC strains occurred 7 days after pre-colonization with B. fragilis 9343 strains. Colonization levels by each strain in each mouse were tracked by collection of faecal pellets over a period of 4 weeks, plating on selective BHIS agar plates (B. fragilis 9343 on BHIS plus erythromycin; B. fragilis 638R and ATCC on BHIS plus tetracycline), and subsequent absolute quantification of c.f.u. by normalization of each sample to the initial pellet weight. Differences in the strain ratio of c.f.u between groups at each time point was assessed using Mann–Whitney *U*-tests. Non-parametric tests were used following Shapiro-Wilk analysis for normality of data at each time point. Mice were confirmed to be sterile before colonization by qPCR with primers targeting the 16S rRNA gene and free of non-Bacteroides contamination by plating faecal pellets on non-selective LB and BHIS plates incubated under either anaerobic or aerobic conditions<sup>43</sup>.

#### Bioinformatic analysis of rAID-1 clusters

The amino acid sequence of the B. fragilis NCTC 9343 polyimmunityassociated XerD-like tyrosine recombinase (BF9343\_RS08045) was used as a query against a custom database of 423 Bacteroidales genomes downloaded from GenBank.rAID clusters in Bacteroidales genomes were identified based on the following criteria: (i) presence of a 5' XerD-like tyrosine recombinase gene encoding a protein with amino acid identity exceeding 44% (corresponding to an E value of  $1 \times 10^{-100}$ ); (ii) two or more co-directionally oriented downstream genes that possessed (iii) a GC content of 41% or lower. The end of the gene cluster was defined as the stop codon of the last co-directionally oriented gene in the cluster with similar GC content. To identify homologues of genes within rAID-clusters, open-reading frames (ORFs) within the clusters were translated and used as tblastn queries against the NCBI non-redundant nucleotide database. Top hits from these searches were often genes in other rAID clusters; therefore, these hits were discarded. The top non-rAID hit from tblastn searches with an E-value threshold of  $1 \times 10^{-30}$  was selected as the closest homologue. rAID cluster genes were assigned to interbacterial immunity gene families via hmm scans with profiles previously described<sup>2</sup> with an E-value cut-off of  $1 \times 10^{-3}$ . rAID cluster genes were also compared via tblastn with 46 Bacteroidales T6SS immunity genes from subtypes  $1-3^{3,8}$  with an E-value cut-off of  $1\times10^{-10}$ . The percentage amino acid identity with homologues was assessed if sequences could be aligned across more than 80% of their length. Motifenrichment analysis was performed on non-coding sequences within a subset of rAID-1 clusters (14 sequences immediately 3' of the recombinase stop codon, and 86 intergenic sequences between rAID-1 ORFs), using MEME Suite 5.0.2 and default settings<sup>44</sup>.

# Heterologous expression of Bacteroides toxin and immunity genes

To assess the ability of cognate immunity or orphan immunity to neutralize the toxicity of a *Bacteroidales* T6SS effector, genes were cloned into *E. coli* expression vectors pScrhab2-V (effectors) and pPSV39-CV (immunity). Immunity genes were fused with the *P. aeruginosa* ribosome binding site from hcp1 during the cloning process $^{45}$ . All cloning steps for effector genes involved growth of *E. coli* on media containing 0.1% glucose to ensure repression of expression. *E. coli* DH5a cells were co-transformed with pSchraB2 and pPSV39 plasmids bearing genes of interest. Overnight cultures were then grown from single co-transformed colonies to stationary phase in LB broth containing 50  $\mu g \, \text{ml}^{-1}$  trimethoprim, 15  $\mu g \, \text{ml}^{-1}$  gentamycin, and glucose. Cells were collected from these cultures and washed to remove glucose before back-dilution to an OD $_{600}$  of 0.05 into LB broth containing 50  $\mu g \, \text{ml}^{-1}$  trimethoprim, 15  $\mu g \, \text{ml}^{-1}$  gentamycin, 0.05% rhamnose, and 1 mM isopropyl  $\beta$ -D-1-thiogalat opyranoside  $^{45,46}$ . Cultures were then grown for 8 h shaking at 37 °C before

plating to allow quantification of c.f.u. Experiments were performed with technical triplicates for each of at least two biological replicates.

### Gene expression analysis of AID-1 and rAID-1 systems of *B. ovatus* 3725

To assess the level of expression of genes in the AID-1 and rAID-1 systems of B. ovatus 3725, bacterial cells were first grown overnight on BHIS blood agar plates containing gentamycin. Cells were then resuspended in BHIS to an  $OD_{600}$  of 3.0 for *B. ovatus* monocultures, or to an  $OD_{600}$  of 0.3 for mixed co-culture experiments with B. fragilis 9343 at tenfold excess (OD<sub>600</sub> of 3.0). Then,  $5 \mu l$  volumes of bacterial mixtures were then spotted on BHIS blood agar plates. Plates were incubated at 37 °C for 2 h under anaerobic culture conditions before cells were collected directly in Buffer RLT plus \(\beta\)-mercaptoethanol (205-ul spots per condition per replicate, Qiagen RNeasy Micro Kit). Two separate rounds of DNase treatment were performed (Qiagen RNase-free DNase, Thermo Fisher Scientific Turbo DNase-free kit). RNA samples were confirmed to be free of genomic DNA by PCR with primers targeting the Bacteroides 16S rRNA gene. cDNA was generated using the High Capacity cDNA Reverse Transcription Kit (Applied Biosciences). Following synthesis, cDNA was diluted 1:10. qPCR (primers listed in Supplementary Table 10) was performed using SSO Universal SYBR Green Supermix (Bio-Rad) on a CFX96 machine (Bio-Rad). Genomic DNA was used to generated standard curves<sup>47</sup>. Differences in gene expression between samples were performed by normalization to the expression level of *B. ovatus* 3725 gyrB. Primers targeting gyrB were designed to target regions of the genes that are highly polymorphic between B. fragilis and B. ovatus, and species-specificity for B. ovatus was confirmed by PCR using B. fragilis genomic DNA48.

#### **Reporting summary**

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

#### **Data availability**

All data required to assess the conclusion of this research are available in the main text and Supplementary Information, have been deposited at the Sequence Read Archive (SRA) under BioProject accession PRJNA484981 or are available from https://github.com/borensteinlab/T6SS).

#### **Code availability**

Python and R scripts used in this work are available for download (https://github.com/borenstein-lab/T6SS).

- Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature 486, 207–214 (2012).
- Qin, J. et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 490, 55–60 (2012).
- Li, J. et al. An integrated catalog of reference genes in the human gut microbiome. Nat. Biotechnol. 32, 834–841 (2014).
- Truong, D. T. et al. MetaPhIAn2 for enhanced metagenomic taxonomic profiling. Nat. Methods 12, 902–903 (2015).
- Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012).

- Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).
- Luo, R. et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1, 18 (2012).
- Besemer, J., Lomsadze, A. & Borodovsky, M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.* 29, 2607–2618 (2001).
- Bacic, M. K. & Smith, C. J. Laboratory maintenance and cultivation of bacteroides species. Curr. Protoc. Microbiol. 9, 13C.1.1–13C.1.21 (2008).
- Koropatkin, N. M., Martens, E. C., Gordon, J. I. & Smith, T. J. Starch catabolism by a prominent human gut symbiont is directed by the recognition of amylose helices. Structure 16, 1105–1115 (2008).
- Degnan, P. H., Barry, N. A., Mok, K. C., Taga, M. E. & Goodman, A. L. Human gut microbes use multiple transporters to distinguish vitamin B<sub>12</sub> analogs and compete in the gut. Cell Host Microbe 15, 47–57 (2014).
- 38. Hoffman, L. R. et al. *Escherichia coli* dysbiosis correlates with gastrointestinal dysfunction in children with cystic fibrosis. *Clin. Infect. Dis.* **58**, 396–399 (2014).
- Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLOS Comput. Biol. 13, e1005595 (2017).
- Martens, E. C., Chiang, H. C. & Gordon, J. I. Mucosal glycan foraging enhances fitness and transmission of a saccharolytic human gut bacterial symbiont. Cell Host Microbe 4, 447-457 (2008)
- García-Bayona, L. & Comstock, L. E. Bacterial antagonism in host-associated microbial communities. Science 361. eaat2456 (2018).
- Lim, B., Zimmermann, M., Barry, N. A. & Goodman, A. L. Engineered regulatory systems modulate gene expression of human commensals in the gut. Cell 169, 547– 558 (2017).
- Paik, J. et al. Potential for using a hermetically-sealed, positive-pressured isocage system for studies involving germ-free mice outside a flexible-film isolator. Gut Microbes 6, 255–265 (2015).
- Bailey, T. L. et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 37, W202–W208 (2009).
- Silverman, J. M. et al. Haemolysin coregulated protein is an exported receptor and chaperone of type VI secretion substrates. Mol. Cell 51, 584–593 (2013).
- Cardona, S. T. & Valvano, M. A. An expression vector containing a rhamnose-inducible promoter provides tightly regulated gene expression in *Burkholderia cenocepacia*. *Plasmid* 54. 219–228 (2005).
- Bookout, A. L., Cummins, C. L., Mangelsdorf, D. J., Pesola, J. M. & Kramer, M. F. Highthroughput real-time quantitative reverse transcription PCR. Curr. Protoc. Mol. Biol. 73, 15.81–15.8.28 (2006).
- Caro-Quintero, A. & Ochman, H. Assessing the unseen bacterial diversity in microbial communities. Genome Biol. Evol. 7, 3416–3425 (2015).

Acknowledgements We thank the UW GNAC for assistance with gnotobiotic experiments. We thank C. Sears, A. Goodman, T. Kuwahara and E. Martens for providing *Bacteroides* strains. This work was supported by National Institutes of Health (NIH) grants Al080609 (to J.D.M.), P30DK089507 (to L.R.H. as pilot study PI), R01DK095869 (to L.R.H.), K99GM129874 (to B.D.R.), R01GM124312 (to E.B.), and New Innovator Award DP2AT00780201 (to E.B.), and the Burroughs Wellcome Fund (to J.D.M.). A.J.V. was supported by a postdoctoral fellowship from the Natural Sciences and Engineering Research Council of Canada. B.D.R. was supported by a Simons Foundation-sponsored Life Sciences Research Foundation postdoctoral fellowship. E.B. is a Faculty Fellow of the Edmond J. Safra Center for Bioinformatics at Tel Aviv University. J.D.M. is an HHMI Investigator.

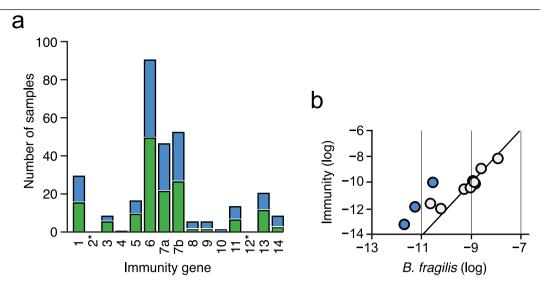
**Author contributions** B.D.R., A.J.V., A.M.H., S.B.P., E.B. and J.D.M. designed the study. B.D.R. and D.T.S. performed in vitro growth experiments; A.J.V., B.D.R. and M.C.R. performed bioinformatic analyses; B.D.R., C.E.P. and L.R.H. isolated and sequenced genomes of gut bacteria; B.D.R., D.T.S., A.M.H. and S.B.P. performed gnotobiotic mouse experiments; B.D.R., A.J.V., M.C.R., D.T.S., A.M.H., S.B.P., E.B. and J.D.M. analysed data; and B.D.R., A.J.V., S.B.P., E.B. and J.D.M. wrote the manuscript.

Competing interests The authors declare no competing interests.

#### Additional information

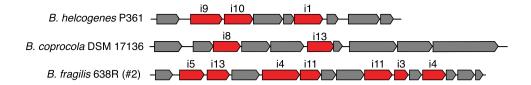
 $\textbf{Supplementary information} \ is \ available \ for \ this \ paper \ at \ https://doi.org/10.1038/s41586-019-1708-z.$ 

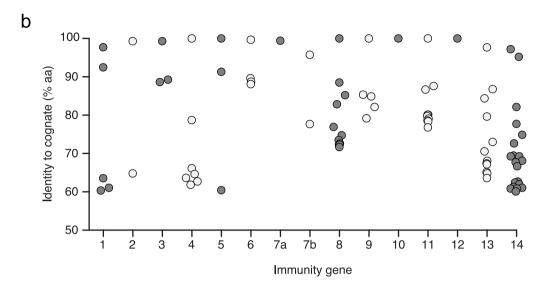
Correspondence and requests for materials should be addressed to E.B. or J.D.M. Peer review information Nature thanks Melanie Blokesch, Kevin Foster and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Reprints and permissions information is available at http://www.nature.com/reprints.



Extended Data Fig. 1 | Prevalence of *B. fragilis*-specific orphan immunity genes in adult and infant microbiomes. a, Number of adult human gut microbiome samples in which the indicated immunity genes (1–14, GA3\_i1–14 from ref.  $^8$ ) can be detected at an 80% nucleotide identity threshold and an abundance more than tenfold that of *B. fragilis* marker genes. Bars coloured as in Fig. 1a, and asterisks indicate immunity genes without orphan representation.

**b**, Comparison of abundance of B. fragilis-specific T6SS immunity genes with B. fragilis species-specific marker genes in infant microbiome samples  $^{16}$  (Supplementary Table 4). Abundances are calculated as in Fig. 1a. Samples in which immunity gene abundance exceeds that of Bacteroides by over tenfold (blue) are highlighted.

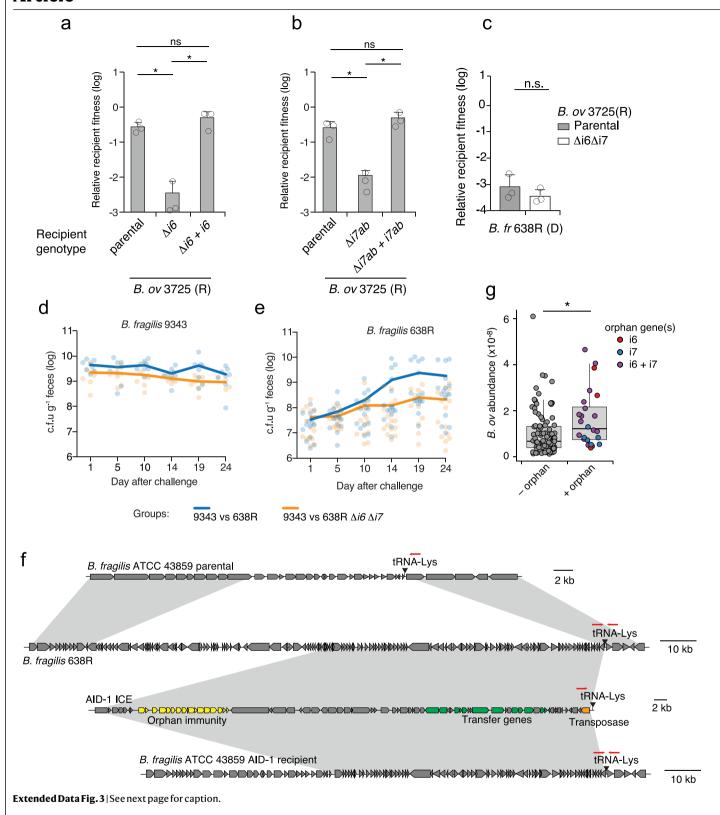




# $Extended \, Data \, Fig. \, 2 \, | \, Diversity \, and \, genomic \, context \, of \, or phan \, immunity \, genes \, in \, human \, gut \, microbiomes \, and \, diverse \, \textit{Bacteroides} \, species.$

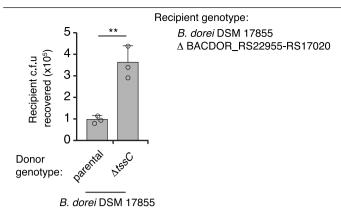
 ${\bf a}, Representative AID-1 gene clusters containing homologues of the indicated {\it B. fragilis} T6S immunity genes from the indicated reference genomes. {\bf b}, Data$ 

points indicate the amino acid identity of unique genes homologous to indicated *B. fragilis*-specific T6SS cognate immunity genes identified through BLAST analysis of the IGC<sup>29</sup> (n=88 genes, maximum  $E=1\times10^{-40}$ ; minimum percentage identity, 60%).

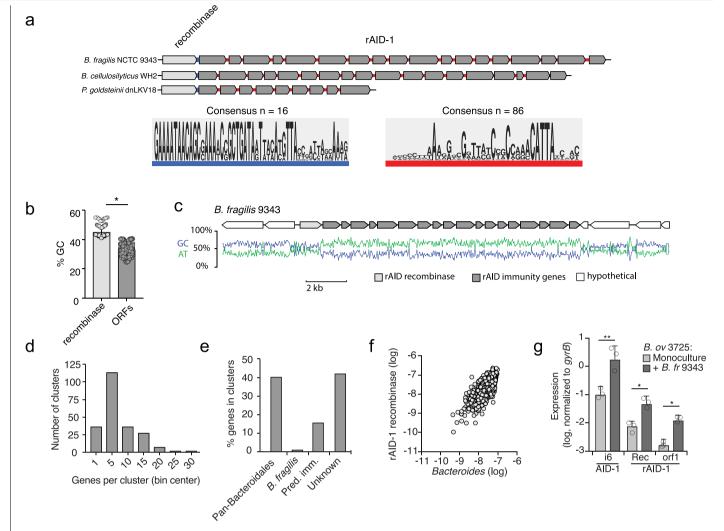


Extended Data Fig. 3 | Orphan immunity genes specifically enhance the fitness of Bacteroides strains in vitro and in vivo. a, b, T6SS-dependent competitiveness of parental strains of B. ovatus 3725 and the indicated mutant and complemented derivatives during in vitro growth competition experiments with B. fragilis 9343. Relative recipient fitness was determined by calculating the ratio of final to initial c.f.u. and normalizing to the corresponding experiment with B. fragilis 9343 lacking tssC (T6S-inactive). Data are mean  $\pm$ s.d. of three independent biological replicates. \*P<0.01, unpaired two-tailed t-test. c, T6SS-dependent competitiveness of a parental strain of B. ovatus 3725 or a strain bearing in-frame deletions of indicated orphan immunity genes, during in vitro growth competition experiments with an orthogonal effector-bearing B. fragilis 638R parental strain or a derivative strain lacking tssC (T6S-inactive). Relative recipient fitness and statistics were calculated as in a and b. n = 3 independent biological replicates. d, e, Recovery of B. fragilis 9343 (d) or 638R (e) and the

indicated orphan immunity mutant derivative from pairwise competitions of the strains in germ-free mice. Lines indicate the mean at each time point (n=8 mice per group for each of two independent experiments). Alternating time points of these data are included in ratio form in Fig. 3c.  $\mathbf{f}$ , Schematic depicting genomic loci for the B. fragilis ATCC 43859 parental strain, the B. fragilis 638R AID-1 donor strain, the AID-1 system, and the ATCC 43859 AID-1 recipient. Grey shading indicates homology; red arrows indicate the position of PCR primers used to infer insertion of the AID-1 element at the tRNA<sup>1</sup>/<sub>2</sub>s insertion site.  $\mathbf{g}$ , Abundance of B. ovatus in samples lacking detected orphan immunity genes (–) and samples in which the indicated orphan immunity genes were assigned to B. ovatus (+). Abundances are calculated as in Fig. 1a. \*P<0.001, Wilcoxon ranksum test. n=128 non-orphan samples, n=24 samples containing orphan immunity. For box plots, the middle line denotes the median; the box denotes the interquartile range (IQR); and the whiskers denote 1.5× the IQR.



# **Extended Data Fig. 4** | The GA2 system of Bacteroidales mediates interbacterial antagonism. Recovery of Bacteroides dorei DSM 17855 cells lacking GA2\_e14-i14 (BACDOR\_RS22955-17020) from two-strain in vitro growth competition experiments with the indicated donor strains. n=3 technical replicates representative of three biological replicates. \*\*P<0.01, unpaired two-tailed t-test.



Extended Data Fig. 5 | rAID-1 systems include conserved and repetitive intergenic sequences and bear hallmarks of horizontal gene transfer. a, Left, motif enrichment analysis from the intergenic sequences immediately 3' of the recombinase stop codon to the start codon of the first downstream open reading frame within 16 randomly selected rAID-1 gene clusters. This region is highlighted in blue in three representative rAID-1 systems shown above. Right, motif enrichment analysis from all 86 intergenic sequences between the ORFs of six rAID-1 clusters (B. fragilis NCTC 9343, B. cellulosilyticus WH2, B. ovatus 3725, Paraprevotella clara YIT 11840, Parabacteroides goldsteinii dnLKV18, and Parabacteroides gordonii MS-1)<sup>44</sup>. This region is highlighted in red in three representative rAID-1 systems shown above. b, Average G + C nucleotide content of rAID-1-associated recombinase versus rAID-1 predicted ORFs (n = 226).

\*\*\*\*P < 0.0001, unpaired two-tailed t-test. c, Schematic depicting the G + C and A + T nucleotide content across a representative rAID-1 system from B. fragilis 9343. d, Frequency distribution of gene number in rAID-1 clusters (n = 1,247

genes in 226 clusters). Bin width is five genes. **e**, Composition of genes in rAID-1 clusters (n = 226 clusters) as determined by profile HMM scans and BLAST analysis against a curated database of Bacteroidales T6SS immunity genes<sup>2,8</sup>. **f**, Comparison of the total abundances of rAID-1-associated predicted recombinases and the *Bacteroides* genus in adult microbiome samples derived from the HMP and MetaHIT studies (Supplementary Table 8). Abundance values are calculated as in Fig. 1; genus abundance corresponds to the sum of all *Bacteroides* spp. (calculated individually as the average of species-specific marker gene abundances). **g**, Results of qRT–PCR analyses for the indicated *B. ovatus* 3725 genes belonging to AID-1 (i6, M088\_1971) or AID-1 clusters (Rec, recombinase, M088\_1401; orf1, M088\_1400) under conditions of growth in mono- or co-culture with *B. fragilis* 9343 for 2 h. Data are mean ± s.d. of three independent biological replicates. \*P<0.05, \*\*P<0.01, Wilcoxon two-tailed sign-rank test.



Corresponding author(s):	Mougous
Last updated by author(s):	YYYY-MM-DD

# **Reporting Summary**

X Life sciences

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see <u>Authors & Referees</u> and the <u>Editorial Policy Checklist</u>.

Sto	HUSTICS			
For	all statistical analys	ses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.		
n/a	Confirmed			
	The exact san	nple size $(n)$ for each experimental group/condition, given as a discrete number and unit of measurement		
	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly			
	The statistical test(s) used AND whether they are one- or two-sided  Only common tests should be described solely by name; describe more complex techniques in the Methods section.			
$\boxtimes$	A description of all covariates tested			
	A description	of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons		
	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient)  AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)			
	For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i> ) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted <i>Give P values as exact values whenever suitable.</i>			
$\boxtimes$	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings			
$\times$	For hierarchio	cal and complex designs, identification of the appropriate level for tests and full reporting of outcomes		
$\boxtimes$	Estimates of e	effect sizes (e.g. Cohen's $d$ , Pearson's $r$ ), indicating how they were calculated		
	'	Our web collection on <u>statistics for biologists</u> contains articles on many of the points above.		
So	ftware and o	code		
Poli	cy information abo	out <u>availability of computer code</u>		
		Provide a description of all commercial, open source and custom code used to collect the data in this study, specifying the version used OR state that no software was used.		
Data analysis		The following publicly-available software packages were used to analyze data in this manuscript: MetaPhlAn2.0, bowtie2, samtools, BLAST, SoapDeNovo2, GeneMarkS, Unicycler, HMMERv3.2.1, MEME Suite 5.0.2, GraphPad Prism 7.		
		tom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.		
Da	ta			
All	manuscripts must - Accession codes, ur - A list of figures that	out <u>availability of data</u> include a <u>data availability statement</u> . This statement should provide the following information, where applicable: nique identifiers, or web links for publicly available datasets have associated raw data y restrictions on data availability		
All r	materials generated a	as part of this manuscript, including strains and plasmids, will be made readily available upon request following publication.		
Fie	eld-spec	ific reporting		

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Ecological, evolutionary & environmental sciences

Behavioural & social sciences

# Life sciences study design

All studies must disclose on these points even when the disclosure is negative.			
Sample size	No statistical methods were used to determine sample size for in vitro experiments and experiments involving gnotobiotic mice. For mouse experiments, we used reasonable numbers considering limitations of housing and maintenance under gnotobiotic conditions.		
Data exclusions	No data were excluded from reporting		
Replication	All experiments reported in this study were reproducibly replicated.		
Randomization	For gnotobiotic experiments, mice from multiple litters were randomized into treatment groups when possible.		
Blinding	Investigators were not blinded to group allocation during gnotobiotic experiments		

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems		Methods		
n/a	Involved in the study	n/a	Involved in the study	
$\boxtimes$	Antibodies	$\boxtimes$	ChIP-seq	
$\boxtimes$	Eukaryotic cell lines	$\boxtimes$	Flow cytometry	
$\boxtimes$	Palaeontology	$\boxtimes$	MRI-based neuroimaging	
	Animals and other organisms			
$\boxtimes$	Human research participants			
$\boxtimes$	Clinical data			

#### Animals and other organisms

Policy information about <u>studies involving animals</u>; <u>ARRIVE guidelines</u> recommended for reporting animal research

Laboratory animals

All mice used in this study were adult (8-12 weeks of age) Swiss-Webster females.

Wild animals

Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals

were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.

Field-collected samples

For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.

Ethics oversight The University of Washington Office of Animal Welfare IACUC approved the protocol governing the use of gnotobiotic mice in this study.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

# Altered chromosomal topology drives oncogenic programs in SDH-deficient GISTs

https://doi.org/10.1038/s41586-019-1668-3

Received: 3 December 2017

Accepted: 10 September 2019

Published online: 16 October 2019

William A. Flavahan<sup>1,2,11</sup>, Yotam Drier<sup>1,2,10,11</sup>\*, Sarah E. Johnstone<sup>1,2</sup>, Matthew L. Hemming<sup>3,4</sup>, Daniel R. Tarjan<sup>1,2</sup>, Esmat Hegazi<sup>1,2</sup>, Sarah J. Shareef<sup>1,2</sup>, Nauman M. Javed<sup>1,2</sup>, Chandrajit P. Raut<sup>5</sup>, Benjamin K. Eschle<sup>6</sup>, Prafulla C. Gokhale<sup>6</sup>, Jason L. Hornick<sup>7</sup>, Ewa T. Sicinska<sup>8</sup>, George D. Demetri<sup>3,4,9\*</sup> & Bradley E. Bernstein<sup>1,2,9\*</sup>

Epigenetic aberrations are widespread in cancer, yet the underlying mechanisms and causality remain poorly understood<sup>1-3</sup>. A subset of gastrointestinal stromal tumours (GISTs) lack canonical kinase mutations but instead have succinate dehydrogenase (SDH) deficiency and global DNA hyper-methylation<sup>4,5</sup>. Here, we associate this hypermethylation with changes in genome topology that activate oncogenic programs. To investigate epigenetic alterations systematically, we mapped DNA methylation, CTCF insulators, enhancers, and chromosome topology in KIT-mutant, PDGFRA-mutant and SDH-deficient GISTs. Although these respective subtypes shared similar enhancer landscapes, we identified hundreds of putative insulators where DNA methylation replaced CTCF binding in SDH-deficient GISTs. We focused on a disrupted insulator that normally partitions a core GIST super-enhancer from the FGF4 oncogene. Recurrent loss of this insulator alters locus topology in SDH-deficient GISTs, allowing aberrant physical interaction between enhancer and oncogene. CRISPR-mediated excision of the corresponding CTCF motifs in an SDH-intact GIST model disrupted the boundary between enhancer and oncogene, and strongly upregulated FGF4 expression. We also identified a second recurrent insulator loss event near the KIT oncogene, which is also highly expressed across SDH-deficient GISTs. Finally, we established a patient-derived xenograft (PDX) from an SDH-deficient GIST that faithfully maintains the epigenetics of the parental tumour, including hypermethylation and insulator defects. This PDX model is highly sensitive to FGF receptor (FGFR) inhibition, and more so to combined FGFR and KIT inhibition, validating the functional significance of the underlying epigenetic lesions. Our study reveals how epigenetic alterations can drive oncogenic programs in the absence of canonical kinase mutations, with implications for mechanistic targeting of aberrant pathways in cancers.

The human genome is partitioned into physical domains, often termed topologically associated domains (TADs), by chromosomal boundaries established by the DNA-binding insulator protein CTCF and cohesin<sup>6-9</sup>. Many proto-oncogenes and master regulators are isolated in such domains and thus protected from promiscuous enhancer interactions<sup>10</sup>.

Mutations of tricarboxylic-acid-cycle-related enzymes, including SDH and isocitrate dehydrogenase (IDH), are initiating events in many tumour types<sup>1,4,5</sup>. These lesions cause accumulation of succinate and 2-hydroxyglutarate, respectively, which inhibit demethylases and are associated with DNA hypermethylation and other epigenetic alterations<sup>5,11,12</sup>.

The CTCF insulator is methylation-sensitive and may be displaced by DNA methylation<sup>13-15</sup>. We previously showed that the *PDGFRA* oncogene is aberrantly activated by insulator defects in *IDH*-mutant glioma<sup>16</sup>.

We hypothesized that SDH deficiency alters chromosome topology to drive GIST tumorigenesis (Fig. 1a). GISTs are the most common gastrointestinal tract sarcoma. They are typically caused by gainof-function mutations of the KIT or PDGFRA oncogenes that render these receptor tyrosine kinases (RTKs) active and ligand-independent<sup>17</sup>. However, approximately 15% of GISTs lack these defining mutations, and have instead lost SDH expression due to mutation or transcriptional

Department of Pathology and Center for Cancer Research, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. 2Broad Institute of MIT and Harvard, Cambridge, MA, USA. 3Center for Sarcoma and Bone Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. 4Department of Medical Oncology, Dana-Farber Cancer Institute and Harvard Medical School Boston, Boston, MA, USA. 5Department of Surgery, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. Experimental Therapeutics Core, Belfer Center for Applied Cancer Science, Dana-Farber Cancer Institute, Boston, MA, USA. 7Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. Bepartment of Oncologic Pathology, Dana-Farber Cancer Institute and Haryard Medical School, Boston, MA, USA, Ludwig Center at Haryard, Haryard Medical School, Boston, MA, USA, <sup>10</sup>Present address: The Lautenberg Center for Immunology and Cancer Research, IMRIC, Faculty of Medicine, The Hebrew University, Jerusalem, Israel. <sup>11</sup>These authors contributed equally: William A. Flavahan. Yotam Drier. \*e-mail: yotam.drier@mail.huji.ac.il: George Demetri@dfci.harvard.edu: Bernstein.Bradley@mgh.harvard.edu

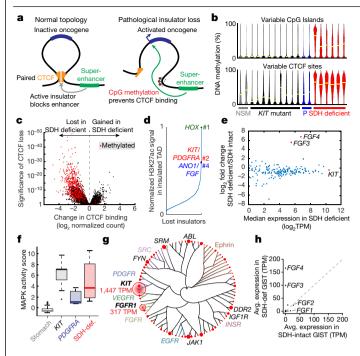


Fig. 1 | Insulator dysfunction in SDH-deficient GISTs. a, Proposed mechanism of epigenetic oncogene activation. Left, oncogene shielded from superenhancer by CTCF insulator, which creates a topological boundary. Right, CTCF insulator displaced by DNA methylation, allowing the super-enhancer to contact and induce the oncogene. b, Violin plots depict DNA methylation levels over the 10,000 most variable CpG island promoters (top) and CTCF sites (bottom) in normal stomach muscle (NSM; n = 2), and KIT mutant (n = 9), PDGFRA mutant (P; n=2) and SDH-deficient GISTs (n=6). Yellow bars indicate mean.  $\mathbf{c}$ , Volcano plot depicts differential CTCF occupancy between SDH-deficient (n=6) and SDH-intact (n=8) GISTs. Sites that gain DNA methylation in SDH-deficient GISTs are indicated in red (>25% increase, two-sided t-test false discovery rate (FDR) <5%). **d**, Plot depicts H3K27ac peaks near lost CTCF insulators (y axis) rank ordered by signal strength.  ${f e}$ , Scatter plot depicts genes (points) separated from a super-enhancer by a CTCF loop anchor that is lost in SDH-deficient GIST. Genes are positioned according to their relative (yaxis) and absolute median expression (x axis) in SDH-deficient GISTs. Potentially deregulated gene targets (outliers) include oncogenes FGF3, FGF4 and KIT (red); see also Supplementary Information. TPM, transcripts per million. f, Box plot depicts average expression of MAPK signature genes in RNA-seq data for normal stomach (n = 262), and KIT mutant (n=10), PDGFRA mutant (n=3) and SDH-deficient GISTs (n=8). Boxes depict 25th, 50th and 75th percentiles, and whiskers depict extreme values. g, Radial phylogenetic tree depicts tyrosine kinase gene expression in SDHdeficient GISTs. Each branch is one tyrosine kinase, arranged by similarity, and with major families depicted by colour. The area of each red circle is proportional to the average expression of the kinase. h, Scatter plot depicts average expression of FGF ligands in SDH-intact (x axis) and SDH-deficient (yaxis) GISTs. FGF3 and FGF4 are highly expressed in SDH-deficient GISTs (bold). For all panels, n values indicate number of biologically independent specimens.

silencing of SDH subunit genes (SDHA-D) $^{18}$ . We collected an initial cohort of clinically defined specimens, including  $11\,KIT$ -mutant,  $2\,PDGFRA$ -mutant and  $8\,SDH$ -deficient tumours (Supplementary Table 1). We used hybrid-selection bisulfite sequencing to profile DNA methylation of over  $160,000\,CTCF$  sites and representative promoters in 17 of these tumours and  $2\,n$  ormal stomach muscle samples (see Methods). Consistent with previous reports  $^5$ , SDH-deficient GISTs exhibited CpG island hypermethylation (Fig. 1b). In addition, a substantial fraction of CTCF sites were methylated in this GIST subtype (Fig. 1b).

We next identified candidate insulators and enhancers in these tumours by mapping CTCF and histone H3 acetylated at lysine 27(H3K27ac) by chromatinimmunoprecipitation and sequencing (ChIPseq). Overall patterns of enhancer acetylation were largely consistent

across GISTs, relative to gastrointestinal carcinomas (Extended Data Fig. 1a). By contrast, comparison of genome-wide CTCF binding profiles revealed that approximately 5% of sites were specifically lost in SDH-deficient GISTs (Fig. 1c). CTCF loss was accompanied by notable increases in DNA methylation at these sites (Fig. 1c and Extended Data Fig. 1b, c). Given that DNA methylation has been established to prevent CTCF binding<sup>13–15</sup>, this suggests that hypermethylation displaces CTCF from hundreds of candidate insulators in SDH-deficient tumours.

To investigate the impact of CTCF loss on genome topology, we used HiC to map TADs and TAD boundaries genome-wide in GIST-T1, a human cell line with an oncogenic KIT mutation and intact SDH expression<sup>19</sup>. We also used HiChIP to map CTCF loops and loop anchors, which correspond to TADs and boundaries, respectively<sup>20</sup> (Extended Data Fig. 1d). We used these maps to predict insulator losses likely to alter topology and gene expression. Of the 1,236 sites that lose CTCF and gain methylation in SDH-deficient GISTs, 688 corresponded to loop anchors. We reasoned that disruption of these loop anchors could alter topology and, in certain cases, permit aberrant enhancer-promoter interactions (Fig. 1a). Therefore, we further curated this list using enhancer maps and expression data. This highlighted 60 CTCF loop anchors that would normally have partitioned a large 'super-enhancer' from a gene, but that were lost in SDH-deficient GISTs (Fig. 1d, e and Supplementary Table 2). Top hits included lost CTCF insulators in the FGF3 and FGF4 locus (chromosome 11q13) and the KIT locus (chromosome 4q12) (Extended Data Fig. 2a, b).

Although SDH-deficient GISTs lack *KIT* or *PDGFRA* mutations<sup>18</sup>, our insulator analysis raised the possibility that RTKs may instead be epigenetically deregulated. This prompted us to examine the expression of RTKs, ligands and downstream signalling programs. First, we found that a signature for mitogen-activated protein kinase (MAPK) targets is highly expressed and suggestive of active RTK signalling in SDH-deficient GISTs (Fig. 1f; see Methods). Second, a systematic analysis of tyrosine kinase gene expression revealed that *KIT* and FGF receptor 1 (*FGFR1*) are the most highly expressed RTKs in SDH-deficient GISTs (Fig. 1g). Third, we found that *FGF3* and *FGF4* were expressed at remarkably high levels, and were both specific to the SDH-deficient subtype (Fig. 1h). *FGF3* and *FGF4* are established oncogenes<sup>21</sup>, and FGF signalling could help to explain the poor efficacy of KIT inhibitors in SDH-deficient GISTs<sup>22,23</sup>. We therefore investigated the mechanisms underlying this striking and specific upregulation of FGF ligands.

FGF3 and FGF4 reside in an approximately 250 kb TAD flanked by boundaries that contain CTCF-binding sites (Fig. 2a). The adjacent TAD on the 11q side contains a large cluster of enhancers or super-enhancer. This super-enhancer overlaps the gene ANO1, which encodes the GIST clinical biomarker also known as DOG-1 ('discovered on GIST-1')<sup>24</sup>. The super-enhancer is highly acetylated and ANO1 is highly expressed in all GIST subtypes (Extended Data Fig. 2a). Notably, the TAD boundary that partitions this super-enhancer from the FGF genes, which we refer to as the 'FGF insulator', contains several CTCF-binding sites (Fig. 2a).

We hypothesized that disruption of CTCF binding could compromise the FGF insulator and allow the *ANOI* super-enhancer to contact and activate the FGF genes. The FGF insulator contains two strong and several weak CTCF-binding sites. Although these sites are consistently bound in *KIT*- and *PDGFRA*-mutant tumours and normal stomach muscle control samples, all five are markedly reduced in SDH-deficient GISTs (Extended Data Fig. 2c). The strongest CTCF-binding site, which is closest to the *ANOI* super-enhancer, is almost completely lost in the SDH-deficient samples (Extended Data Fig. 2d). Consistently, it is methylated specifically in SDH-deficient tumours. This suggests that the FGF insulator has switched to a methylated state that occludes CTCF binding.

To assess the impact of CTCF loss on boundary integrity, we performed circularized chromatin conformation capture sequencing (4C-seq) on four SDH-intact and three SDH-deficient GISTs. We designed a 'viewpoint' primer that enabled us to quantify contacts between a central position in the *ANOI* super-enhancer and other genomic positions at high resolution (Fig. 2b). In SDH-intact tumours and stomach

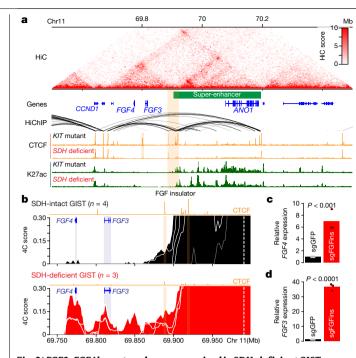


Fig. 2| FGF3-FGF4 locus topology reorganized in SDH-deficient GISTs.

a, Genomic views of the FGF3-FGF4 and ANO1 loci depict baseline chromosome topology (HiC, red), genes (blue), CTCF-CTCF loop interactions (HiChIP, arcs, with darkness indicating significance), CTCF binding (ChIP-seq, orange) and candidate enhancers (H3K27ac ChIP-seq, green). HiC/HiChIP data are for the SDH-intact model GIST-T1, whereas CTCF and H3K27ac data are for representative clinical specimens (see also Extended Data Fig. 2). ANO1 superenhancer (green bar) and FGF insulator (orange shading) are indicated. b, Traces depict 4C-seq interaction frequency (yaxis) between the ANOI super-enhancer viewpoint primer (dashed white line) and genomic positions in the FGF3-FGF4-ANO1 locus (x axis). Data are shown for SDH-intact GISTs (n = 4; top), normal stomach muscle (n=1; grey line, top) and SDH-deficient GISTs (n=3; bottom). CTCF binding profiles for representative SDH-intact (top) and SDHdeficient (bottom) tumours are also shown (orange). Genes (blue) and CTCF sites in the FGF insulator (orange) are highlighted. c, d, Plots depict relative FGF4 (c) and FGF3 (d) expression in GIST-T1 cells expressing CRISPR-Cas9 and control sgRNA (black) or sgRNAs targeting the two CTCF sites in the FGF insulator (red). Bars indicate mean of three biologically independent replicates (dots). Pvalues by two-sided t-test.

muscle control samples, we detected robust interactions throughout the ANO1 TAD, but not beyond its boundaries, consistent with robust FGF insulator function. In SDH-deficient tumours, however, the same super-enhancer viewpoint physically interacted with sequences well beyond the boundary, including with the FGF3 and FGF4 genes, which are ~200 kb from the viewpoint (Fig. 2b and Extended Data Fig. 3a-c). These data suggest that FGF3-FGF4 locus topology is profoundly altered in SDH-deficient GISTs, with CTCF insulator loss allowing aberrant contacts between the ANO1 super-enhancer and FGF ligand genes.

To assess directly whether FGF insulator loss affects FGF gene expression, we used genome editing to disrupt the insulator in GIST-T1 cells, which harbour a GIST-like enhancer landscape and retain CTCF binding and boundary function. We used CRISPR-Cas9 and short guide RNAs (sgRNAs) to edit the motifs underlying the two strongest CTCF sites in the FGF insulator (Extended Data Fig. 3d). This resulted in a sixfold induction of FGF4, and a 35-fold induction of FGF3 (Fig. 2c, d). These data directly link insulator loss to the marked upregulation of FGF ligands in SDH-deficient GISTs.

Notably, a switch between CTCF-bound and DNA-methylated insulator states underlies genomic imprinting<sup>13,14</sup>. FGF insulator loss might therefore also represent a stable epigenetic event or 'epimutation' that

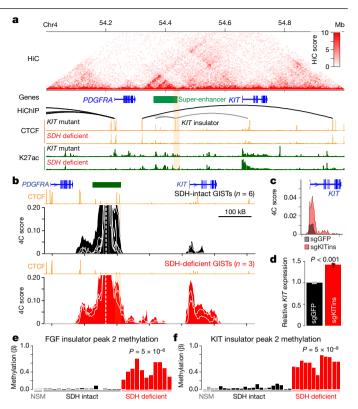
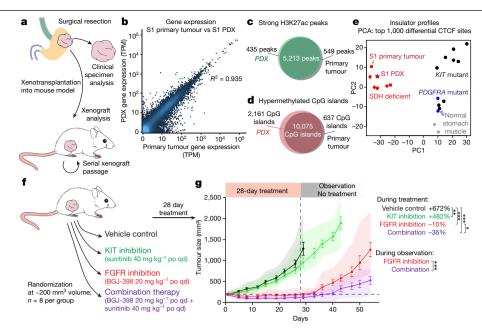


Fig. 3 | KIT-PDGFRA locus topology reorganized in SDH-deficient GISTs.

a, Genomic views of PDGFRA and KIT loci depict baseline chromosome topology (HiC, red), genes (blue), CTCF-CTCF loop interactions (HiChIP, arcs), CTCF binding (ChIP-seq, orange) and candidate enhancers (H3K27ac ChIP-seq, green). HiC/HiChIP data are for the SDH-intact GIST model GIST-T1, whereas CTCF and H3K27ac data are for representative clinical specimens (see also Extended Data Fig. 2). KIT super-enhancer (green bar) and KIT insulator (orange shading) are indicated, **b**. Traces depict 4C-seq interaction frequency (vaxis) between the KIT super-enhancer viewpoint primer (dashed white line) and genomic positions in the KIT-PDGFRA locus (x axis). Data are shown for SDHintact GISTs (n = 6, top) and SDH-deficient GISTs (n = 3, bottom). CTCF profiles for representative SDH-intact (top) and SDH-deficient (bottom) tumours are also shown. Genes (blue) and CTCF-binding sites in the KIT insulator (orange) are highlighted.c, Traces depict 4C-seq interaction signal between the KIT superenhancer viewpoint primer and the KIT gene in GIST-T1 cells expressing Cas9 and control (black) or KIT insulator targeting sgRNAs (red). d, Plot depicts relative KIT expression in GIST-T1 cells expressing Cas9 and control (black) or KIT insulator targeting sgRNAs (red). Bar indicates mean of three biologically independent replicates (dots). P values by two-sided t-test. e, f, FGF and KIT insulator fraction methylation (β-value) evaluated in an expanded cohort of GIST tumours by locus-specific bisulfite sequencing. e, Bar plot depicts average  $methylation \, levels \, across \, six \, CpGs \, within \, FGF \, insulator \, CTCF \, peak \, 2 \, in \, normal \,$ stomach muscle (NSM: n=2). SDH-intact GISTs (n=17) and SDH-deficient GISTs (n=11). f, Bar plot depicts average methylation levels across four CpGs within KIT insulator CTCF peak 2 in normal stomach muscle (n=2), SDH-intact GISTs (n=20) and SDH-deficient GISTs (n=12) (n values indicate number ofbiologically independent tumours).

effects a single allele. In five of our SDH-deficient samples, we identified heterozygous single nucleotide polymorphisms (SNPs) within an FGF3 or FGF4 exon. In four of these cases, analysis of the informative SNP in RNA-seq data revealed that the FGF ligand gene was mono-allelically expressed (Extended Data Fig. 4a-c). By contrast, ANO1 was bi-allelically expressed, suggesting that the biased FGF expression reflected allelespecific insulator loss. Consistently, in one SDH-deficient tumour with a heterozygous SNP near the CTCF site, we confirmed that only one allele of the FGF insulator was methylated (Extended Data Fig. 4d). In a second tumour with an informative SNP near the ANO1 super-enhancer



**Fig. 4** | **SDH-deficient GIST PDX trial confirms dependence on FGF signalling. a**, Specimen collection and generation of PDX model. **b**, Scatter plot compares expression of genes (points) between primary tumour S1(*x* axis) and PDX (*y* axis) per RNA-seq. Pearson correlation is indicated. **c**, Venn diagram depicts overlap between strong H3K27ac ChIP-seq peaks in primary tumour (black) and PDX (green). **d**, Venn diagram depicts overlap between hypermethylated CpG islands in primary tumour (black) and PDX (red) per bisulfite sequencing. **e**, Scatter plot depicts principal component analysis (PCA) on the top 1,000 differential CTCF sites for primary tumours (coloured by subtype) and PDX (star). PDX and originating tumour (S1) both cluster with SDH-deficient GISTs

(red). **f**, Experimental design of the pre-clinical trial. Following xenograft implantation and growth, mice were randomized to four treatment groups and treated with the indicated regimen daily for 28 days (oral daily). Observation was continued until the clinical endpoint (tumour volume of 2,000 mm³). **g**, Plot depicts tumour volume during treatment and observation periods. Points represent mean tumour volume, error bars represent s.e.m., and shading represents the range of tumour volumes for n=8 biologically independent xenografts per group. Relative tumour volume immediately following treatment cessation (day 29) is indicated at the right of the panel. P values reflect the difference in tumour growth between group per two-way ANOVA (P < 0.05, P < 0.001, P < 0.0001).

4C-seq viewpoint primer, we confirmed that the aberrant interaction between super-enhancer and FGF4 was also strongly biased to one allele (Extended Data Fig. 4e, f). These data suggest that insulator loss, topological reorganization and FGF induction reflect a stable epigenetic alteration propagated in the malignant clone.

In addition to the FGF insulator, our screen identified a top-ranked CTCF insulator loss in the *KIT* locus. This hit was of interest given that *KIT* is an established GIST oncogene, and given prior reports of crosstalk between FGF and KIT signalling  $^{22,23}$ . HiC and HiChIP data reveal that the *KIT* gene is contained within a -600 kb TAD (Fig. 3a). This large TAD contains within it a smaller insulated domain (-100 kb) that is flanked by CTCF sites. This smaller domain harbours a large super-enhancer that is highly acetylated in all GIST specimens examined (Extended Data Fig. 2b). It is partitioned from KIT by a topological boundary that we refer to as the KIT insulator.

The KIT insulator contains two strong CTCF sites separated by around 7 kb. Both sites gain methylation and lose CTCF binding in SDH-deficient GISTs (Extended Data Fig. 2e, f). To determine whether CTCF loss is associated with altered KIT locus topology, we performed 4C-seq using a viewpoint primer in the insulated super-enhancer (Fig. 3b). In SDH-intact tumours, the super-enhancer engages in robust interactions throughout the small insulated domain, but not beyond its boundaries (Extended Data Fig. 5a, b). In SDH-deficient tumours, however, the super-enhancer interacts with sequences well beyond the KIT insulator (Extended Data Fig. 5c, d), consistent with loss of CTCF binding and boundary function. Notably, quantification of interaction signals in SDH-deficient tumours indicated that approximately 15–20% of interactions made by this superenhancer viewpoint are with the KIT promoter and gene, compared with 1–5% in SDH-intact tumours (Extended Data Fig. 5d).

To test directly whether CTCF loss alters *KIT* locus topology, we edited the two CTCF motifs in the KIT insulator in GIST-T1 cells (Extended Data Fig. 5e) and evaluated locus topology by 4C-seq. Although the KIT

insulator boundary was clearly evident in control GIST-T1 cells (Fig. 3a), it was compromised in the edited cells, as demonstrated by frequent contacts between super-enhancer and KIT (Fig. 3c). We also considered the impact of insulator loss on *KIT* expression. Although GIST-T1 cells already highly express a constitutively active form of this oncogene, we found that insulator disruption further increased *KIT* expression by around 50% under culture conditions that partially mimic SDH deficiency (Fig. 3d and Extended Data Fig. 6). Although this proportional increase is modest, it corresponds to a substantial increase in transcriptional output given high baseline *KIT* expression in GIST-T1 cells.

Our hypothesis that FGF and KIT insulator losses drive SDH-deficient GIST predicts that these insulators should be recurrently disabled, and the corresponding oncogenes consistently expressed across tumours. We therefore examined insulator methylation across 32 GIST specimens from our original cohort and an additional validation cohort. Both insulators were highly methylated in all SDH-deficient cases, but not in any SDH-intact tumours or normal controls (Fig. 3e, f). Consistently, CTCF binding to these insulators was compromised in all six SDH-deficient GISTs evaluated, but retained in all SDH-intact tumours and normal stomach muscle controls (Extended Data Fig. 2c-f). Furthermore, these CTCF sites were consistently unmethylated across multiple non-malignant cell and tissue types, including a population enriched for interstitial cells of Cajal (ICCs), the presumed GIST cell of origin (Extended Data Fig. 7a). Finally, FGF4 is consistently expressed across SDH-deficient tumours yet it is only expressed at very low or undetectable levels in KIT-mutant GISTs, PDGFRA-mutant GISTs and ICCs (Fig. 1h and Extended Data Fig. 7b). The recurrence and specificity of these insulator losses support their functional significance in SDH-deficient GISTs.

Finally, we evaluated directly whether signalling through FGFR and/ or KIT is required for tumour growth. Although we are unaware of any in vitro SDH-deficient GIST models, we successfully established an in vivo PDX model from one of our SDH-deficient GIST specimens (KIT and PDGFRA wild type) (Fig. 4a). Model and parental tumour have remarkably similar RNA expression. H3K27acenhancer landscapes, methylation and CTCF binding profiles (Fig. 4b-e). The PDX also maintains characteristic enhancers and CTCF insulator losses in the FGF and KIT loci, and strongly expresses FGF3. FGF4 and KIT (Extended Data Fig. 8a.b). These data support the fidelity of this SDH-deficient GIST model.

We therefore tested the efficacy of FGFR and KIT inhibitors in this model. We used BGJ-398, a potent and selective inhibitor of FGFR1-4 in clinical development<sup>25</sup>, and sunitinib, a drug approved for GIST with potent activity against unmutated KIT<sup>26</sup>. We dosed PDX mice for 28 days with BGJ-398 (20 mg kg<sup>-1</sup>), sunitinib (40 mg kg<sup>-1</sup>) or a combination of the two(Fig.4f). Single agent sunitinib minimally suppressed tumour growth, consistent with the drug-resistant phenotype of the SDH-deficient GIST subtype and prior reports that cross-talk between FGF and KIT signalling confers resistance to KIT inhibition<sup>22,23</sup>. By contrast, single agent BGJ-398 completely suppressed tumour growth throughout the dosing period, strongly supporting a critical role for FGF signalling in tumorigenesis (Fig. 4g and Supplementary Table 3). Sensitivity to FGFR inhibition is specific to this GIST subtype as BGJ-398 lacks efficacy against SDH-intact PDX models<sup>23</sup>. Notably, the combination of FGFR and KIT inhibitors resulted in the most durable response, with growth suppression well beyond the dosing period (Fig. 4g and Extended Data Fig. 8c, d). These pre-clinical data indicate that both RTK signalling pathways drive SDH-deficient GIST, and strongly support the significance of the underlying epigenetic lesions.

In conclusion, we identify multiple epigenetic lesions that converge to activate RTK signalling and proliferation in SDH-deficient GIST. We show that the characteristic hypermethylation in these tumours is associated with pervasive insulator losses, topological reorganization of the FGF and KIT loci, and particularly potent induction of the FGF4 and FGF3 oncogenes. Although our data do not address the precise cellular contexts in which these lesions arise, it is notable that KIT signalling regulates proliferation of the presumed GIST cell of origin, ICC<sup>27</sup>. Similarly, the ANO1 gene, the super-enhancer of which aberrantly drives FGF3 and FGF4 expression, encodes an ion channel that is highly expressed and essential for ICC<sup>28,29</sup>. Hence, topological changes that deregulate FGF and KIT expression could lead to unchecked signalling in these precursors. Although the corresponding loci are genetically wild type in SDH-deficient GIST, the functional significance of their deregulation is supported by the prevalence of gain-of-function KIT mutations in SDHintact GISTs and by a recent report that FGF4 is genetically amplified in a rare subset of KIT/PDGFRA/SDH/RAS quadruple wild-type GIST<sup>17,30</sup>. Our pre-clinical PDX data substantiate their significance and establish proof of concept for the rapeutic intervention. Given that few stable epigenetic events have been established as drivers of tumorigenesis<sup>2</sup>, our nomination of two novel functional lesions represents an important advance.

#### Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1668-3.

- Kaelin, W. G. Jr & McKnight, S. L. Influence of metabolism on epigenetics and disease Cell 153, 56-69 (2013).
- Jones, P. A. & Baylin, S. B. The epigenomics of cancer. Cell 128, 683-692 (2007).
- Flavahan, W. A., Gaskell, E. & Bernstein, B. E. Epigenetic plasticity and the hallmarks of cancer. Science 357, eaal2380 (2017).
- Janeway, K. A. et al. Defects in succinate dehydrogenase in gastrointestinal stromal tumors lacking KIT and PDGFRA mutations. Proc. Natl Acad. Sci. USA 108, 314-318
- Killian, J. K. et al. Succinate dehydrogenase mutation underlies global epigenomic divergence in gastrointestinal stromal tumor. Cancer Discov. 3, 648-657 (2013).
- Bickmore, W. A. & van Steensel, B. Genome architecture; domain organization of interphase chromosomes. Cell 152, 1270-1284 (2013).
- Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions, Nature 485, 376-380 (2012).
- Dekker, J. & Mirny, L. The 3D genome as moderator of chromosomal communication. Cell 164, 1110-1121 (2016).
- Rao, S. S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping, Cell 159, 1665-1680 (2014).
- Hnisz, D. et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods, Science 351, 1454-1458 (2016).
- Xiao, M. et al. Inhibition of a-KG-dependent histone and DNA demethylases by fumarate and succinate that are accumulated in mutations of FH and SDH tumor suppressors Genes Dev. 26, 1326-1338 (2012).
- Lu, C. et al. IDH mutation impairs histone demethylation and results in a block to cell differentiation. Nature 483, 474-478 (2012).
- Bell, A. C. & Felsenfeld, G. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. Nature 405, 482-485 (2000).
- Hark, A. T. et al. CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. Nature 405, 486-489 (2000).
- Liu, X. S. et al. Editing DNA methylation in the mammalian genome. Cell 167, 233-247.e17 (2016).
- 16. Flavahan, W. A. et al. Insulator dysfunction and oncogene activation in IDH mutant gliomas, Nature 529, 110-114 (2016)
- Hirota, S. et al. Gain-of-function mutations of c-kit in human gastrointestinal stromal tumors. Science 279, 577-580 (1998).
- Boikos, S. A. & Stratakis, C. A. The genetic landscape of gastrointestinal stromal tumor lacking KIT and PDGFRA mutations. Endocrine 47, 401-408 (2014).
- Taguchi, T. et al. Conventional and molecular cytogenetic characterization of a new human cell line, GIST-T1, established from gastrointestinal stromal tumor. Lab. Invest. 82, 663-665 (2002).
- Mumbach, M. R. et al. HiChIP: efficient and sensitive analysis of protein-directed genome 20. architecture, Nat. Methods 13, 919-922 (2016).
- Arao, T. et al. FGF3/FGF4 amplification and multiple lung metastases in responders to sorafenib in henatocellular carcinoma. Henatology 57, 1407-1415 (2013).
- Javidi-Sharifi, N. et al. Crosstalk between KIT and FGFR3 promotes gastrointestinal stromal tumor cell growth and drug resistance, Cancer Res. 75, 880-891 (2015)
- Li, F. et al. FGFR-mediated reactivation of MAPK signaling attenuates antitumor effects of imatinib in gastrointestinal stromal tumors. Cancer Discov. 5, 438-451 (2015)
- West, R. B. et al. The novel marker, DOG1, is expressed ubiquitously in gastrointestinal stromal tumors irrespective of KIT or PDGFRA mutation status. Am. J. Pathol. 165, 107-113 (2004).
- 25. Pal, S. K. et al. Efficacy of BGJ398, a fibroblast growth factor receptor 1-3 inhibitor, in patients with previously treated advanced urothelial carcinoma with FGFR3 alterations. Cancer Discov. 8, 812-821 (2018).
- Janeway, K. A. et al. Sunitinib treatment in pediatric patients with advanced GIST following failure of imatinib, Pediatr, Blood Cancer 52, 767-771 (2009)
- Sircar, K. et al. Interstitial cells of Cajal as precursors of gastrointestinal stromal tumors. Am. J. Surg. Pathol. 23, 377-389 (1999).
- Gomez-Pinilla, P. J. et al. Ano1 is a selective marker of interstitial cells of Cajal in the human and mouse gastrointestinal tract. Am. J. Physiol. Gastrointest. Liver Physiol. 296, G1370-G1381 (2009).
- Singh, R. D. et al. Ano1, a Ca2+-activated Cl1 channel, coordinates contractility in mouse intestine by Ca2+ transient coordination between interstitial cells of Caial. J. Physiol. 592. 4051-4068 (2014)
- Urbini, M. et al. Gain of FGF4 is a frequent event in KIT/PDGFRA/SDH/RAS-P WT GIST. Genes Chromosom, Cancer 58, 636-642 (2019).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

#### Methods

#### Primary GIST specimens and cell culture models

Epigenetically characterized clinical samples were obtained as frozen specimens from Brigham and Women's Hospital or from the Massachusetts General Hospital Pathology Tissue Bank. The validation cohort was obtained as FFPE samples from the BWH tissue bank. All samples were acquired with Institutional Review Board approval, and were de-identified before receipt. *PDGFRA* and *KIT* mutational status were confirmed through Sanger sequencing for frozen specimens, while SDH status was determined by immunohistochemistry (details below).

Samples were also examined via RNA-seq and ChIP-seq input controls (details below) in order to look for mutations or copy number changes in all FGF ligand and receptor genes—no copy number alterations were found and no sequence variants were detected other than known annotated SNPs.

The GIST-T1 cell line was obtained from Cosmo Biosciences <sup>19</sup>. Cells were passaged in DMEM with 10% serum,  $1\times$  antibiotics and  $1\times$  Glutamax (Life Technologies). For pseudohypoxia experiments, cells were treated with  $200~\mu\text{M}$  deferoxamine mesylate (Sigma) or vehicle control (water) for 72~h. For succinate conditions, cells were cultured in  $20~\mu\text{M}$  dimethyl-succinate (Sigma), which was slowly added to a cell culture dish containing standard media and allowed to dissolve before addition of cells.

#### Chromatin immunoprecipitation

ChIP-seq was performed as described previously. In brief, cultured cells or minced frozen tissue were crosslinked in 1% formaldehyde and snap frozen in liquid nitrogen before storage at -80 °C for at least overnight. Sonication of samples were calibrated such that DNA was sheared to between 300 and 700 bp fragment length. CTCF was precipitated with a monoclonal rabbit CTCF antibody, clone D31H2 (Cell Signaling no. 3418). Histone H3K27 acetylation was immunoprecipitated with antibody from Active Motif (no. 39133). ChIP DNA was used to generate sequencing libraries by end repair (End-It DNA repair kit, Epicentre), 3' A base overhang addition via Klenow fragment (NEB) and ligation of barcoded sequencing adapters. Barcoded fragments were amplified via PCR. Libraries were sequenced as 38-bp end reads on an Illumina NextSeq500 instrument. Processed genomic data has been deposited into GEO under accession number GSE107447, while raw sequencing data has been deposited into dbGaP (phs001906.v1.p1). See also Supplementary Table 4.

Reads were aligned to the reference genome (hg38) using BWA aln version 0.7.4<sup>31</sup>, removing reads with mapping quality score <10. For H3K27 acetylation ChIP-seq and input controls, PCR duplicates were removed by Picard toolkit 2.9.2. Peaks were called with HOMER 4.932 against input controls. To call all H3K27ac peaks, we used 'histone' settings. To call super-enhancers<sup>33</sup>, we used 'super' settings and no local filtering. CTCF peaks were called with 'factor' settings. To measure H3K27ac correlations, signal at the union of the peaks (5 kb window around the centre) was calculated by featureCounts 1.6.2<sup>34</sup>. We downloaded and reprocessed publicly available data of other gastrointestinal tumours for comparison (GSM196964535, GSM196965735, GSM2058055<sup>36</sup>, GSM2058056<sup>36</sup>, GSM2131266<sup>37</sup> and GSM2131280<sup>37</sup>). The dendrogram is based on unweighted average distance linkage of the Pearson correlations between the 10,000 most variable peaks, although analysis results were similar when comparing correlations over all peaks.

#### Hybrid selection bisulfite sequencing

Hybrid selection probes were designed to capture -160,000 CTCF-binding sites, and -5,000 promoters. CTCF bind sites lists were collated from ENCODE (as downloaded from UCSC genome browser, table wgEncodeRegTfbsClusteredV3, Release 4) as well as additional CTCF maps of primary cholangiocarcinoma and glioma  $^{16}$ . Total genomic DNA was isolated using the DNAeasy Blood & Tissue Kit (Qiagen) and sheared

using the Covaris LE220. Ampure XP beads (Agencourt) were used to size select gDNA fragments within 150-320 bp and sheared distribution was verified via BioAnalyzer (Agilient). One microgram of gDNA was end repaired, 3' A base tailed (KAPA Hyper Prep Kit no. KK8502) and ligated to sequencing adaptor (Roche SeqCap Epi Enrichment System). Ligated products were purified using Ampure XP beads. Following bead clean-up, products were bisulfite-converted using the EZ DNA Methylation-Lightning Kit (Zymo Research) and then PCR amplified using KAPA HiFi U+ HotStart ReadyMix (KAPA no. KK2800). Equal concentrations of each library were then combined in sets of three or four along with SeqCap Epi universal and indexing oligos and bisulfite capture enhancer (SeqCap Epi Accessory Kit). Each pool was lyophilized using TOMY Micro-Vac (MV100), resuspended in hybridization buffer (SeqCap Epi Hybridization and Wash Kit), and then hybridized to SeqCap Epi Probe Pool (Roche) for 72 h at 47 °C in a thermocycler. Following the 72 h incubation, captured bisulfite-converted libraries were recovered (SeqCap Pure Capture Bead Kit) at 47 °C in a thermocycler for 45 min, with intermediate vortexing. Capture beads were washed (SeqCap Hybridization and Wash Kit) in a 47 °C water bath. Captured bisulfite-converted libraries were then amplified via PCR (SeqCap Epi Accessory Kit). Libraries were sequenced with 10% PhiX spike-in as 100-bp end reads on the HiSeq2500 in rapid run mode.

Hybrid-selection bisulfite sequencing (HSBS) data were processed by methylCtools  $0.9.4^{38}$ , using BWA mem version 0.7.12, and aligned to human reference hg38. Owing to the sizes of the captured fragments, probe capture resulted in an effective coverage of about 600 bp around CTCF sites. PCR duplicates were removed by Picard toolkit 2.9.2. DNA methylation levels were called by methylCtools 0.9.4 in loci covered by at least five reads. Methylation at 36,281 CTCF-binding sites that are bound in GIST tumours and covered by the assay were used for downstream analysis.

#### **HiC and HiChIP**

In situ HiC was performed as described<sup>9</sup>. CTCF HiChIP was performed as described<sup>20</sup>. In brief, 3 tubes of ~5 million GIST-T1 cells were crosslinked in 1% formaldehyde (2 replicates for HiChIP, 1 for HiC). For HiChIP, cells were lysed using HiC lysis buffer as described. Chromatin was digested with 375 U Mbol restriction enzyme (NEB, R0147). After heat inactivation of restriction enzyme and marking of ends with biotin-14-dATP (Life Technologies, 19524-016), DNA was ligated using T4 buffer (NEB. B0202). Chromatin was sheared by Covaris LE220. Chromatin immunoprecipitation was performed with 30 µl of monoclonal rabbit CTCF antibody, clone D31H2 (Cell Signaling 3418). Protein was bound by Protein G beads and after washing was incubated in DNA elution buffer. Eluant was treated with proteinase, crosslinks were reversed and the sample was Zymo purified (Zymo DNA Clean and Concentrator D4003). Biotinylated DNA was pulled down with M280 Streptavidin beads (Invitrogen 11205D) and the DNA was fragmented with Tn5 and libraries were constructed with Nextera kit (Illumina). For HiC, cells were lysed in HiC lysis buffer and chromatin was digested with 200 U Mbol (NEB, RO147) overnight. Ends were marked and DNA was ligated as in HiChIP. DNA was precipitated and sheared by Covaris E220. Biotinylated DNA was pulled down by T1 Streptavidin beads (Life Technologies, 65602). End-repair, A-tailing and adaptor ligation were performed as described<sup>9</sup>. Libraries were prepared using Phusion High-Fidelity DNA Polymerase (NEB, M0530). HiChIP and HiC libraries were sequenced as 75-bp end reads on an Illumina NextSeq500 instrument. Data were processed using HiC-Pro<sup>39</sup> and visualized by the WashU EpiGenome Browser<sup>40</sup> and the R package Sushi41.

The two HiChIP replicates showed high similarity, and therefore were merged for the rest of the analysis. CTCF–CTCF loops were called from HiChIP data with hichipper  $0.7.3^{42}$ . Only loops with FDR <5% and supported by at least 5 reads were considered for downstream analysis.

#### 4C analysis

4C analysis was performed using methods adapted from published protocols<sup>43</sup>. In brief, ~10 million cells from culture or frozen minced tumour specimens were crosslinked in 2% formaldehyde. Fixed samples were lysed in lysis buffer containing protease inhibitor cocktail and mechanically disrupted using a Biomasher tissue grinder (Kimble Chase). Lysis was confirmed using methyl green-pyronin staining. Following lysis, nuclei were digested with NIaIII (NEB) overnight at 37 °C in a thermomixer set to 950 rpm. After heat inactivation of restriction enzyme, diluted nuclei were ligated using T4 DNA ligase (NEB) overnight at 16 °C, followed by RNase and protein as eK treatment. Isolated DNA was then digested overnight in Csp6I (Thermo) at 37 °C, diluted and ligated overnight at 16 °C in order to circularize fragments. Efficacy of each ligation and digestion step was verified via agarose gel electrophoresis. Purified circularized DNA was used as input in PCR reactions to create sequencing libraries. Sixteen reactions per sample were performed, each using 200 ng of circularized 4C DNA (3.2 µg total) in 50 µl using Q5 high-fidelity PCR mastermix (NEB). Primers contained sequencing adapters and barcodes, and annealing sections were as follows: KIT enhancer viewpoint primer: 5'-TTTCTATTTGCTCGTTCATG-3'; KIT nonviewpoint primer: 5'-GGAAACTTCCAAAGTAGGCT-3'; ANO1 enhancer viewpoint primer: 5'-ATGTCGCCCTCCTGCATG-3'; ANO1 non-viewpoint primer: 5'-AGACAAATGAGGCCTGGACG-3'; ANO1 viewpoint primer for SNP detection: 5'-CTCAAACAGACACTCACATG-3'; ANO1 non-viewpoint primer for SNP detection: 5'-TCTTTTTGGTTGGATTGTAGGAGT-3'. Standard 4C sequencing libraries were sequenced as 38-bp end reads on an Illumina NextSeq500, although only the first read (the viewpoint primer read) was used for further processing. 4C sequencing libraries for detecting SNPs were read as 75-bp end reads on the same machine, and the second read was used for SNP detection. Data were analysed via 4Cseqpipe<sup>44</sup>, and median normalized data with a main trend resolution of 22.5 kb were visualized in R.

#### RNA-seq

Total RNA was isolated from clinical GIST samples using the RNeasy Plus Kit (Qiagen) and quality was determined via Bioanalyzer (Agilent). Libraries were prepared using the TruSeq Stranded mRNA Library Prep Kit (Illumina), and equimolar multiplexed libraries were sequenced with single-end 75 bp reads on an Illumina NextSeq 500.

Reads were aligned using STAR 2.5.3<sup>45</sup> to the human reference (hg19). RNA-seq data for SDH intact GISTs were previously published<sup>46</sup>. Gene expression was estimated by featureCounts 1.6.2<sup>34</sup>. TPM values were calculated for these data sets<sup>47</sup>. RNA-seq TPM values for normal stomach was downloaded from GTEx v7<sup>48</sup>.

#### Statistical analysis and reproducibility

CTCF peaks of all GISTs were merged by bedtools merge 2.26<sup>49</sup>, including only peaks with a score >50 and in the top 50,000 as reported by HOMER. Peaks were then centred around CTCF motif where found by FIMO (MEME 4.7)<sup>50</sup> at a 100 bp window around the peak centre, based on JASPAR 2014 CTCF motif MA0139.1 bases 4-19<sup>51</sup>. If multiple motifs were detected, we kept the one with the highest score. Reads were counted by HTSeq 0.6.152. CTCF profiles were normalized by copy number estimates and across samples by standard median ratio. Copy number values were estimated from input by CNVnator 0.353, with 5 kb bins. CTCF sites bound in all samples were used for median ratio normalization as implemented by DESeq2<sup>54</sup>, where a site is considered bound in any given sample if its signal is at least 0.25 of the median of the top 10,000 sites for that sample. Normalization factors were used to scale CTCF signal for visualization (Figs. 2, 4) and differential analysis. Differential CTCF binding was called by DESeq2<sup>54</sup>, identifying 2,106 CTCF-binding sites that significantly lose CTCF binding in our cohort (FDR < 5%, fold change > 2). To estimate methylation at lost insulators sites we measured average methylation over a 250 bp window around the peak centre. We identified 4,502 sites that significantly gain methylation (FDR < 5% as determined by t-test, methylation increase >25%). Sites that both significantly lose CTCF binding and gain DNA methylation were considered 'perturbed' for downstream analysis. We focused on CTCF–CTCF loops that overlap with a perturbed CTCF site and either (1) the loop contains a promoter that is insulated from a super-enhancer (<500 kb away) by the perturbed CTCF site; or (2) the loop contains a super-enhancer (at least half of the super-enhancer resides in the loop) that is insulated from a promoter (<500 kb away) by the perturbed CTCF site. Here we only considered super-enhancers that scored in at least two SDH-deficient GISTs. This resulted in the identification of 60 putatively lost insulators with loss of CTCF and >50% methylation gain, and 167 putatively deregulated genes with >1 TPM median expression across SDH-deficient GIST, 25 of which with >100 TPM (Fig. 1e and Supplementary Table 2).

To test correlations between methylation and CTCF binding we focused only on peaks detected in at least one GIST sample and with an annotated CTCF motif (see above). To empirically estimate the null distribution of the correlation coefficient, CTCF-binding sites were permuted 100 times (Extended Data Fig. 1c).

To derive GIST MAPK activity score, we used previously identified MAPK biomarkers<sup>55</sup>, and published data of Imatinib treatment of a GIST line<sup>56</sup>. We picked biomarkers that were downregulated by the Imatinib treatment (P<0.01by t-test, fold change>2), and expressed in all primary GISTs (>5 TPM). This yielded four biomarkers: DUSP6, ETV5, SPRY2 and SPRY4. Final MAPK score was computed by summing z-scores of the four genes and dividing by 2, as suggested<sup>55</sup>.

For Figs. 2a and 3a, representative ChIP–seq traces were selected from ChIP–seq profiles for the 11 *KIT* mutant and 6 SDH-deficient GISTs characterized, all of which displayed similar results (see Supplementary Table 1 and Extended Data Fig. 2).

For all CRISPR insulator deletions, viral introduction of CRISPR–Cas9+sgRNA vector was repeated three times with separate viral preparations and infections to generate biologically independent replicates.

For epigenomic and transcriptomic characterization of clinical tissue (for example, ChIP-seq, 4C, RNA-seq), multiple clinical specimens were analysed, but technical replicates could not be performed on individual samples due to limited availability of material.

For analysis of clinical tissue, no statistical methods were used to predetermine sample size; rather, all available SDH-deficient tumour specimens with validated SDH loss and enough material available for analysis were tested. For mouse studies, no specific statistical calculations were performed; rather, sample size was determined based on prior experience with similar PDX trials. The investigators were not blinded to allocation during experiments and outcome assessment.

#### **Immunohistochemistry**

Immunohistochemistry was performed on 4- $\mu$ m-thick paraffin-embedded tissue sections using a mouse anti-SDHB monoclonal antibody (clone 21A11AE7;1:200 dilution; Abcam), a rabbit anti-PDGFRA monoclonal antibody (clone D13C6;1:300 dilution; Cell Signaling Technology), and a rabbit anti-KIT polyclonal antibody (1:150 dilution; Dako). Pressure cooker antigen retrieval in citrate buffer (pH 6.1; Dako Target Retrieval Solution) was used for PDGFRA and SDHB. Dako Envision+ secondary antibody was used. The sections were developed using 3,3'-diamin-obenzidine as substrate and counterstained with Mayer's haematoxylin.

#### CRISPR-Cas9 insulator disruption

The following CRISPR sgRNAs were cloned into the LentiCRISPR vector<sup>57</sup>: sgGFP 5′-GAGCTGGACGGCGACGTAAA-3′; sgKIT\_CTCFpeakl 5′-GTCTCTCTCTGCCAGCAGG-3′; sgKIT\_CTCFpeak2 5′-GACTTCC CTGACACTAGATG-3′; sgFGF\_CTCFpeakl 5′-GTCCCACTGCCACC ACAAGA-3′; sgFGF\_CTCFpeak25′-GGGCCAGGCCCGCCGCAGG-3′; sgSD HB 5′-GTGTCTCTTTCAGGCATCTG-3′. sgRNAs were designed to either the GG PAM in the consensus CTCF motifs for CTCF disruption, or to a PAM near the 5′ splice junction of exon 4 of the *SDHB* gene. GIST-T1 cells

were infected with the relevant lentivirus(es) for 48 h. Cells were then selected in 2 µg ml<sup>-1</sup> puromycin for 4 days, with puromycin-containing media refreshed every 2 days. Cells were allowed to recover from puromycin for 1 week before analysis. Genomic DNA was then isolated and the region of interest was amplified using primers with sequencing adaptors and the following annealing regions: *KIT* CTCF Peak1 Forward 5′-TTTGGG ATTCGAGTGACCAC-3′; *KIT* CTCF Peak1 Reverse 5′-TCAGGGCTCAACAG CTTCA-3′; *KIT* CTCF Peak2 Forward 5′-GGAAATAACCTCAACCGGTG-3′; *KIT* CTCF Peak2 Reverse 5′-GACTCGGTCTTGCTCCTCTAA-3′. Libraries were sequenced on an Illumina NextSeq500 as 38 bp end reads, and analysed for editing efficiency. Crosslinked cells were also harvested for ChIP analysis to verify loss of CTCF binding.

#### Quantitative real-time polymerase chain reaction

Total RNA was isolated from GIST-T1 cells using the RNeasy minikit (Qiagen) and used to synthesize cDNA with the SuperScriptIII system (Invitrogen). cDNA was analysed using the SYBR mastermix (Applied  $Biosystems) \, on \, a \, 7500 \, Fast \, Real \, Time \, system \, (Applied \, Biosystems). \, Gene$ expression primers were as follows: KIT forward 5'-GCACAATGGCACGG TTGAAT-3'; KIT reverse 5'-GGTGTGGGGATGGATTTGCT-3'; KITLG forward5'-AGCGCTGCCTTTCCTTATGA-3'; KITLG reverse5'-CCGGGGAC ATATTTGAGGGT-3'; EPAS1 forward 5'-CCACCAGCTTCACTCTCC-3'; EPAS1 reverse 5'-TCAGAAAAGGCCACTGCTT-3'; FGF4 set 1 forward 5'-CCAACAACTACAACGCCTACGA-3';FGF4set1reverse5'-CCCTTCTTGG TCTTCCCATTCT-3'; FGF4 set 2 forward 5'-GCAGCAAGGCCAAGCTCT AT-3';FGF4set2reverse5'-CGGTTCCCCTTCTTGGTCTT-3';FGF3forward 5'-ATGCTTCGGAGCACTACAGC-3'; FGF3 reverse 5'-CACGTACCACAG TCTCTCGG-3'. All gene expression results were normalized to primers for ribosomal protein, large, PO (RPLPO) as follows: forward 5'-TCCC ACTTGCTGAAAAGGTCA-3'; reverse 5'-CCGACTCTTCCTTGGCTTCA-3'.

#### Tyrosine kinome tree visualization

Tyrosine kinase phylogeny data were downloaded from kinase.com<sup>58</sup>. Phylogenetic tree was visualized using the R package ggtree<sup>59</sup>. Expression data of each tyrosine kinase were averaged across the SDH-deficient GISTs, and then plotted on the tree, with the area of the red circles corresponding to the average TPM value in SDH-deficient GISTs.

# Interrogation of public normal tissue, GIST and ICC expression data $\,$

Data for normal tissue expression was obtained from ENCODE $^{60}$ . Mouse interstitial cell of Cajal expression data were previously processed and published $^{61}$ . Raw Affymetrix microarray data (CEL files) of human ICC and GIST samples were downloaded from GEO, under accessions GSE56670 $^{62}$ , GSE77839 $^{63}$ , GSE17743 $^{64}$  and GSE20708 $^{65}$ . CEL files were imported, normalized, and RMA values exported using the R/Bioconductor package affy $^{66}$ .

#### Flow cytometry enrichment and analysis of ICCs

Fresh benign stomach muscle tissue was obtained from the MGH Pathology Tissue Bank and dissected from the gastric epithelium. Tissue was initially manually mechanically dissociated with a sterile scalpel, and then subjected to fine mechanical dissociation through three cycles of 1 min each in a Miltenyi gentleMACS dissociator, resulting in a single-cell suspension. A small portion was removed from the cell suspension to serve as the unlabelled and unpermeabilized control to set size gates and test viability. The remainder of the cell suspension was then incubated in a permeabilizing flow cytometry buffer and stained with ANO1-Alexa488 (Santa Cruz Biotechnology clone C-5), KIT-PE (Biolegend Clone 104D2) or CD45-APC (BD Biosciences clone 2D1) for 30 min at 4 °C. Non-permeabilized control cells were treated with propidium iodide immediately before analysis. Stained cells were analysed and collected on a Sony SH800S cell sorter. Compensation parameters were determined using single-labelled UltraComp eBeads (ThermoFisher). Approximately 1.5 million cells were sorted, of which

 $2,000 \, were \, collected \, as \, CD45^-KIT^+ANO1^+ (ICC \, enriched). \, Cells \, were lysed in a \, small \, volume \, of \, TAE/DTT \, and \, treated \, with \, Proteinase \, K. \, Genomic \, DNA \, in \, the \, lysed \, cell \, mixture \, was \, then \, bisulfite \, converted \, using \, the \, EZ \, DNA \, Methylation \, Gold \, Kit \, (Zymo), \, subjected \, to \, locus \, PCR, \, and \, then \, sequenced \, on \, an \, Illumina \, NextSeq500.$ 

#### PDX model generation and efficacy studies

The PDX model was generated from surgical resection tissue from an SDH-deficient GIST patient who consented to research use of material under an IRB-approved protocol. The surgical sample was implanted subcutaneously in female NSG mice and allowed to grow. Tumour growth was monitored by caliper measurements. Once tumours grew to a size of 1,000 mm<sup>3</sup>, tumours were isolated and cut into pieces of approximately 3 × 3 × 3 mm, dipped in Matrigel (Corning Life Science) and transplanted subcutaneously in additional NSG mice. Tumours were passaged for no more than 10 times. Tumour samples from all passages were banked by viably freezing in Bambanker freezing media (Fisher Scientific) and used for further studies. For efficacy studies, tumour fragments were implanted into 8-week-old NSG mice. Tumours were allowed to establish to  $192 \pm 35.7$  mm<sup>3</sup> in size before randomization into various treatment groups with n = 8/group as: vehicle control (0.1M citrate buffer, pH 4.5), 40 mg kg<sup>-1</sup> sunitinib (LC Laboratories, 0.1M citrate buffer, pH 4.5), 20 mg kg<sup>-1</sup> BGJ-398 (LC Laboratories, acetate buffer, pH 4.6 and PEG300 in 1:1 ratio) or the combination of BGJ-398 and sunitinib. Mice were treated orally once daily for 28 days with these agents. In the BGJ-398 treatment group, 4 of 8 mice, and in the combination treatment group, 7 of 8 mice, lost >15% body weight requiring drug holidays (1-3 days of drug holidays in the single agent BGJ-398 group and 1-15 days of drug holidays in the combination group). Mice were re-started on treatment once body weight recovered to at least >85% of initial body weight. Treatment groups were censored when the tumour volume reached the maximum permissible size of 2,000 mm<sup>3</sup> in any single mouse in that group. Statistics were determined by two-way ANOVA.

All relevant ethical regulations regarding research in animal models were followed. All animal experiments and study protocols were approved by the Dana Farber Cancer Institute Institutional Animal Care and Usage Committee (IACUC). The endpoint criteria for mice were if the total tumour burden/size reaches 2 cm in any direction or tumour volume exceeds 2,000 mm³, and/or if the tumour mass interferes with basic/vital bodily functions or becomes persistently ulcerated, and these criteria were followed for all mice in the study.

#### **Reporting summary**

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

#### **Data availability**

Sequencing data that support the findings of this study have been deposited in GEO with the accession code GSE107447.

- Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760 (2009).
- Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell 38, 576-589 (2010).
- Whyte, W. A. et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell 153, 307–319 (2013).
- Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930 (2014).
- Ooi, W. F. et al. Epigenomic profiling of primary gastric adenocarcinoma reveals superenhancer heterogeneity. Nat. Commun. 7, 12983 (2016).
- Cohen, A. J. et al. Hotspots of aberrant enhancer activity punctuate the colorectal cancer epigenome. Nat. Commun. 8, 14400 (2017).
- McDonald, O. G. et al. Epigenomic reprogramming during pancreatic cancer progression links anabolic glucose metabolism to distant metastasis. Nat. Genet. 49, 367–376 (2017).
- Hovestadt, V. et al. Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing. Nature 510, 537–541 (2014).

- Servant, N. et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. 16, 259 (2015).
- Zhou, X. et al. Exploring long-range genome interactions using the WashU Epigenome Browser. Nat. Methods 10, 375–376 (2013).
- Phanstiel, D. H. Sushi: Tools for visualizing genomics data. R package version 1.16.0. https://www.bioconductor.org/packages/release/bioc/html/Sushi.html (2019).
- Lareau, C. A. & Aryee, M. J. hichipper: a preprocessing pipeline for calling DNA loops from HiChIP data. Nat. Methods 15, 155–156 (2018).
- Splinter, E., de Wit, E., van de Werken, H. J., Klous, P. & de Laat, W. Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: from fixation to computation. *Methods* 58, 221–230 (2012).
- van de Werken, H. J. et al. Robust 4C-seq data analysis to screen for regulatory DNA interactions. Nat. Methods 9, 969–972 (2012).
- 45. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15-21 (2013).
- Hemming, M. L. et al. Gastrointestinal stromal tumor enhancers support a transcription factor network predictive of clinical outcome. *Proc. Natl Acad. Sci. USA* 115, E5746–E5755 (2018).
- Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* 131, 281–285 (2012).
- 48. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
- Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010).
- Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. Bioinformatics 27. 1017–1018 (2011).
- Mathelier, A. et al. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 42, D142–D147 (2014).
- 52. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
- Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 21, 974–984 (2011).
- Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550 (2014).
- Wagle, M. C. et al. A transcriptional MAPK pathway activity score (MPAS) is a clinically relevant biomarker in multiple cancer types. NPJ Precis. Oncol. 7, 2 (2018). Correct publication information for this reference?
- Chi, P. et al. ETV1 is a lineage survival factor that cooperates with KIT in gastrointestinal stromal tumours. *Nature* 467, 849–853 (2010).
- Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. Science 339, 819–823 (2013).
- Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. The protein kinase complement of the human genome. Science 298, 1912–1934 (2002).
- Yu, G. et al. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol. Evol. 8, 28–36 (2017).
- ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).
- Lee, M. Y. et al. Transcriptome of interstitial cells of Cajal reveals unique and selective gene signatures. PLoS One 12, e0176031 (2017).
- Killian, J. K. et al. Recurrent epimutation of SDHC in gastrointestinal stromal tumors. Sci. Transl. Med. 6, 268ra177 (2014).

- Tang, C. M. et al. Hedgehog pathway dysregulation contributes to the pathogenesis of human gastrointestinal stromal tumors via GLI-mediated activation of KIT expression. Oncotarget 7, 78226–78241 (2016).
- Ostrowski, J. et al. Functional features of gene expression profiles differentiating gastrointestinal stromal tumours according to KIT mutations and expression. BMC Cancer 9, 413 (2009).
- Astolfi, A. et al. A molecular portrait of gastrointestinal stromal tumors: an integrative analysis of gene expression profiling and high-resolution genomic copy number. *Lab. Invest.* 90, 1285–1294 (2010).
- Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20, 307–315 (2004).
- Li, Z. et al. Hypoxia-inducible factors regulate tumorigenic capacity of glioma stem cells. Cancer Cell 15, 501–513 (2009).
- Murakami, A. et al. Hypoxia increases gefitinib-resistant lung cancer stem cells through the activation of insulin-like growth factor 1 receptor. PLoS One 9, e86459 (2014).
- Bosch-Marce, M. et al. Effects of aging and hypoxia-inducible factor-1 activity on angiogenic cell mobilization and recovery of perfusion after limb ischemia. Circ. Res. 101, 1310–1318 (2007).

Acknowledgements We thank S. Gillespie, M. Miri, F. Najm, P. van Galen and E. Choy for assistance with clinical samples and analysis, H. Gu and A. Gnirke for sequencing assistance, and E. Gaskell for discussions. W.A.F. is supported by an F32 from the National Cancer Institute. Y.D. is supported by the Tosteson Postdoctoral Fellowship. B.E.B. is the Bernard and Mildred Kayden Endowed MGH Research Institute Chair and an American Cancer Society Research Professor. This research was supported by the National Cancer Institute, the NIH Common Fund. the Starr Cancer Consortium. and the Ludwig Center at Harvard.

Author contributions Conception and experimental design: W.A.F., Y.D., S.E.J., M.L.H., E.T.S., G.D.D. and B.E.B. Methodology and data acquisition: W.A.F., Y.D., S.E.J., M.L.H., D.R.T., E.H., S.J.S., N.M.J., C.P.R., B.K.E., P.C.G., J.L.H., E.T.S., G.D.D. and B.E.B. Analysis and interpretation of data: W.A.F., Y.D., M.L.H., S.J.S., N.M.J., G.D.D. and B.E.B. Manuscript writing and revision: W.A.F., Y.D. and B.E.B.

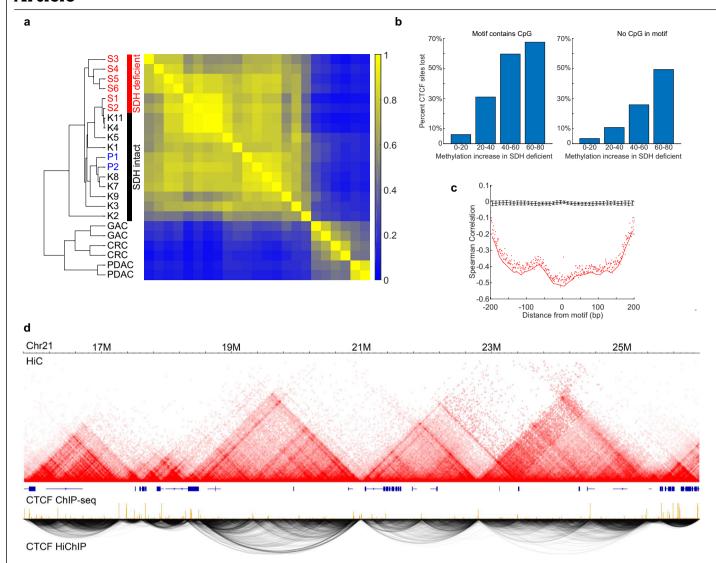
Competing interests B.E.B. is an advisor and equity holder for Fulcrum Therapeutics, 1CellBio, HiFiBio and Arsenal Biosciences, is an advisor for Cell Signaling Technologies, and has equity in Nohla Therapeutics. G.D.D. reports relationships with Novartis, Bayer, Pfizer, EMD-Serono, Sanofi, Ignyta, Roche, Loxo Oncology, AbbVie, Mirati Therapeutics, Epizyme, Daiichi-Sankyo, WIRB Copernicus Group, ZioPharm, Polaris Pharmaceuticals, M.J. Hennessey / OncLive, Adaptimmune, GlaxoSmithKline, Blueprint Medicines, Merrimack Pharmaceuticals, G1 Therapeutics, CARIS Life Sciences, Bessor Pharmaceuticals, ERASCA Pharmaceuticals, CHAMPIONS Oncology, Janssen, PharmaMar; in addition, G.D.D. has a Use patent on imatinib for GIST. licensed to Novartis, with royalties paid to the Dana-Farber Cancer Institute.

#### Additional information

 $\textbf{Supplementary information} \ is available for this paper at \ https://doi.org/10.1038/s41586-019-1668-3.$ 

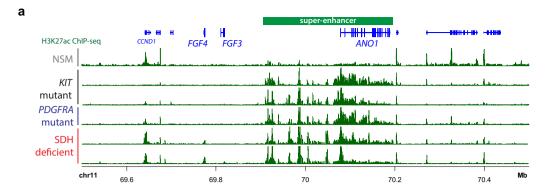
Correspondence and requests for materials should be addressed to Y.D., G.D.D. or B.E.B. Peer review information *Nature* thanks Christian Frezza, Michael Heinrich, Michael Rehli and Peter Scacheri for their contribution to the peer review of this work.

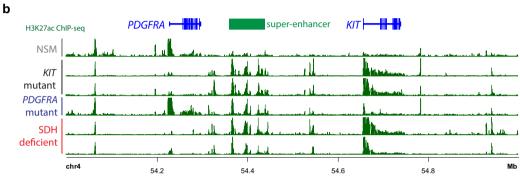
 $\textbf{Reprints and permissions information} \ is \ available \ at \ http://www.nature.com/reprints.$ 

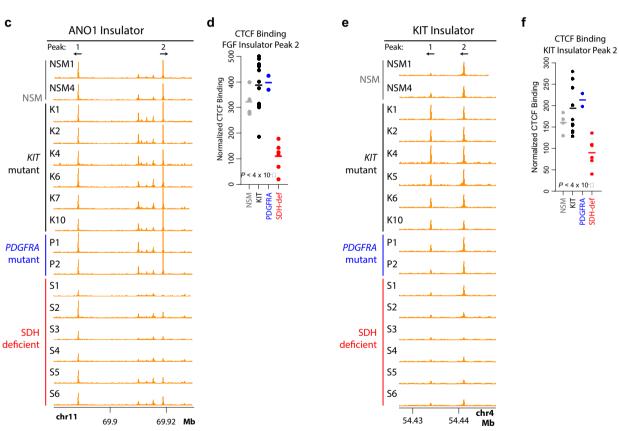


Extended Data Fig. 1 | Epigenomic characterization of GIST. a, ChIP-seq profiles for H3K27ac were compared for GIST specimens and other gastrointestinal tract tumour specimens (GAC, gastric adenocarcinoma; CRC, colorectal cancer; PDAC, pancreatic ductal adenocarcinoma). Heat map depicts pairwise Pearson correlations between the top 10,000 most variable peaks (yellow indicates high correlation; blue indicates low correlation). The dendrogram (left) was derived by unweighted average distance linkage. Enhancer patterns are relatively consistent across GIST subtypes, compared to other tumour types. b, DNA methylation levels in the vicinity of CTCF sites were profiled genome-wide by hybrid-selection bisulfite sequencing. CTCF sites are binned according to the amount their methylation increased in SDH-deficient GISTs, relative to SDH-intact GISTs (methylation change computed over a 250 bp window centred on the motif). For each bin, bar graphs depict the percentage of sites that lose CTCF binding in SDH-deficient GISTs, per ChIP-seq. Separate plots are shown for CTCF sites for which motifs do or do not contain a CpG.

Increased methylation over CTCF sites is associated with more frequent loss of CTCF binding, even when the CTCF motif lacks a CpG.  $\mathbf c$ , Plot depicts correlation between CTCF occupancy and DNA methylation in SDH-deficient GISTs. Red points show Spearman correlations between CTCF ChIP–seq signal and methylation of CpGs at indicated positions relative to the centre of the CTCF motif. Red line reflects correlation to average methylation over 10 bp windows. Randomly permuted data (black) are shown for comparison. Anti-correlation between CTCF occupancy and methylation is evident over a -250-bp binding footprint.  $\mathbf d$ , Genomic views of a representative 10 Mb region on chromosome 21 depict chromosome topology (HiC, red), CTCF binding (ChIP–seq, orange) and CTCF–CTCF loop interactions (HiChIP, black) for the SDH-intact GIST model, GIST-T1. TADs are visible as triangles of enhanced interaction in HiC data, flanked by boundaries that correspond to loop interactions in HiChIP data. Genes (blue) are also indicated.



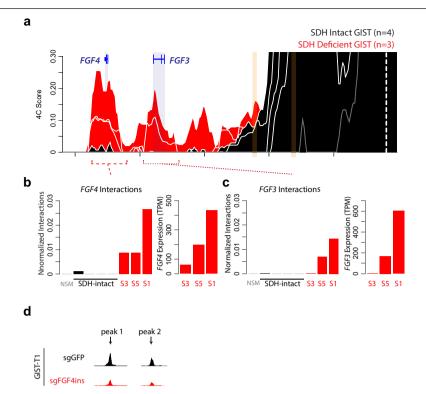




**Extended Data Fig. 2** | See next page for caption.

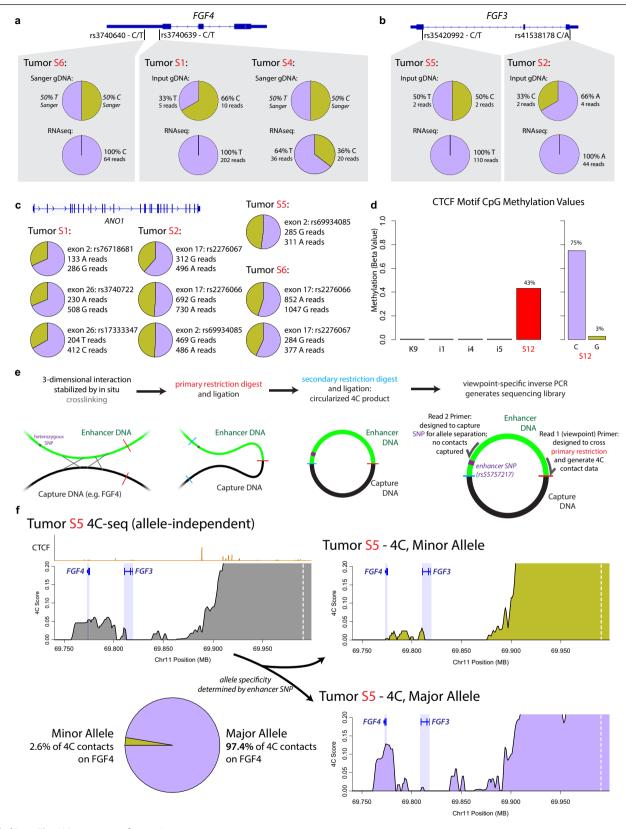
**Extended Data Fig. 2** | **Super-enhancers and insulators in GIST. a**, Traces depict H3K27ac ChIP–seq signal for normal stomach muscle (NSM) and GISTs of the indicated subtype over the *FGF–ANO1* locus. **b**, Traces depict H3K27ac ChIP–seq signal for NSM and GISTs of the indicated subtype over the *PDGFRA–KIT* locus. Genes are indicated in blue, and super-enhancer locations are indicated by green bars. For **a**, **b**, traces are representative of 11 *KIT*-mutant and 6 SDH-deficient tumours with similar results. **c**, Traces depict CTCF binding over the FGF insulator in normal stomach muscle (NSM) and GIST clinical specimens. **d**, Plot depicts CTCF ChIP–seq signal over the strongest CTCF peak in the FGF

insulator in normal stomach muscle (NSM, n=4), and KIT mutant (n=11), PDGFRA mutant (n=2) and SDH-deficient GISTs (n=6). **e**, Traces depict CTCF binding over the KIT insulator in normal stomach muscle (NSM) and GIST clinical specimens. **f**, Plot depicts CTCF ChIP—seq signal over the strongest CTCF peak in the KIT insulator in normal stomach muscle (NSM, n=4), and KIT mutant (n=11), PDGFRA mutant (n=2) and SDH-deficient GISTs (n=6). For **d** and **f**, horizontal bars reflect mean values and P values indicate significance of CTCF loss in SDH-deficient GIST, as determined by the Walt test (via DEseq2<sup>54</sup>). All n values represent the number of biologically independent clinical specimens.



**Extended Data Fig. 3** | **FGF locus 4C-seq data and insulator deletion. a**, Traces depict 4C-seq data at FGF locus, as in Fig. 2b, except graphed on the same axis to allow for direct comparison. **b**, **c**, Bar plots quantify 4C-seq interactions between the super-enhancer viewpoint and FGF4 (**b**) or FGF3 (**c**). Expression of these

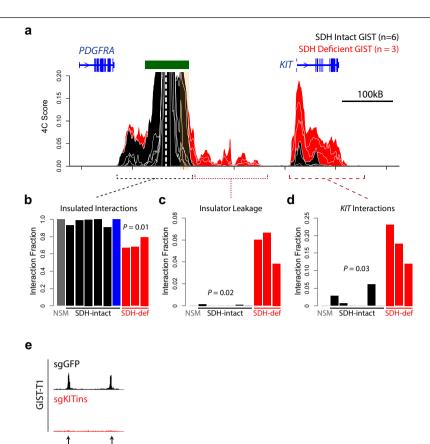
genes in the corresponding SDH-deficient GIST specimens is also shown.  $\mathbf{d}$ , Traces depict CTCF ChIP–seq signal in GIST-T1 cells infected with CRISPR–Cas9 and either a control sgRNA directed at GFP (black, top) or sgRNAs directed against the two indicated CTCF motifs in the FGF insulator (second row, red).



 $\textbf{Extended Data Fig. 4} \, | \, \textbf{See next page for caption}.$ 

#### Extended Data Fig. 4 | Allelic imbalance in FGF3 and FGF4 activation.

 $\textbf{a}, Two\,heterozygous\,SNPs\,in\,\textit{FGF4}\,(both\,3'\,UTR)\,enabled\,us\,to\,evaluate\,allelic$ expression in three SDH-deficient GISTs (tumours S6, S1 and S4). Both alleles for each SNP were detected in DNA sequencing data for these tumours, but only one allele was detected in RNA-seq data of tumours S1 and S6, indicative of monoallelic FGF4 expression. Both alleles are detected in tumour S4, indicating bi-allelic expression of FGF4. **b**, Heterozygous SNPs in FGF3 exons (both synonymous base substitutions) enabled us to evaluate allelic expression in the SDH-deficient GISTs (tumours S2 and S5). In both cases, DNA sequencing confirmed heterozygosity at the genome level (C/A and T/C, respectively), but RNA-seq data demonstrated mono-allelic FGF3 expression. c, Both alleles of heterozygous SNPs in ANO1 exons were found in the RNA-seq data derived from SDH-deficient GIST samples, confirming bi-allelic expression of ANO1. Similarly, both alleles of heterozygous SNPs were found in the histone H3K27ac ChIP-seq data, confirming the bi-allelic nature of the super-enhancer (not shown). d, One SDH-deficient GIST sample was heterozygous for a SNP (rs386829467) located  $about\,50\,bp\,from\,the\,CTCF\,motif\,of\,Peak\,2\,in\,the\,FGF\,insulator.\,Allele-agnostic$ methylation data confirmed 43% methylation of the CTCF peak in this tumour, while essentially no methylation was detected in the SDH-intact tumours (left). Separation of the two alleles using the heterozygous SNP revealed strong allelic bias in the SDH-deficient tumour: one allele was largely unmethylated (~3% methylation), while the other was highly methylated (~75% methylation), consistent with mono-allelic methylation of the CTCF site (right). e, Schematic depicts 4C-seq experimental protocol and primer design for detecting SNPs. DNA elements in close physical proximity are crosslinked and restricted with an enzyme that leaves nucleotide overhangs. These overhangs are then proximity ligated to crosslinked fragments. A second restriction enzyme (with different restriction sites) is then used to circularize the ligated fragments, allowing for  $inverse\,PCR.\,Here\,we selected\,restriction\,enzymes\,and\,designed\,a\,custom\,read\,2$ primer to capture a heterozygous SNP within the super-enhancer. This second read is normally non-informative as contact frequencies are determined through the viewpoint primer (read 1), but in this case enabled us to detect the SNP and assign each ligated fragment to a specific allele. **f**, The left trace (grey) depicts standard 4C-seq data (allele agnostic), which demonstrates strong interaction between super-enhancer viewpoint and FGF4. However, the SNP covered in the non-viewpoint read enabled us to distinguish interactions involving the minor (top right) or major (bottom right) allele. This revealed that  $the \,major\, allele\, (purple)\, is\, responsible\, for\, \text{-}97\%\, of\, super-enhancer-FGF4$ interactions.

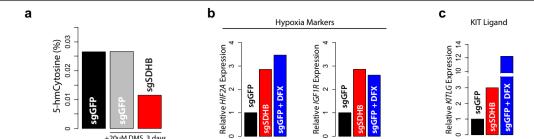


**Extended Data Fig. 5** | *KIT* **locus 4C-seq data and insulator deletion. a**, Traces depict KIT locus 4C-seq data, as in Fig. 3b, except graphed on the same axis to allow for direct comparison.  $\mathbf{b} - \mathbf{d}$ , Bar plots quantify 4C-seq interactions (top, reproduced from Fig. 3b) between the super-enhancer viewpoint and positions within the super-enhancer TAD ( $\mathbf{b}$ ), sequences just beyond the KIT insulator ( $\mathbf{c}$ ),

peak 1

peak 2

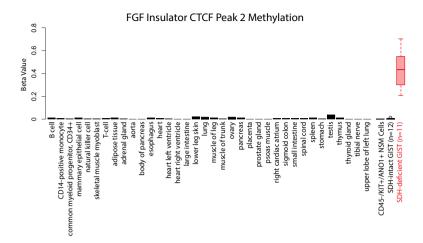
or the  $\it KIT$  gene itself (d).  $\it P$  values indicate significance of difference between SDH-intact and SDH-deficient, by two-sided  $\it t$ -test. e, Traces depict CTCF ChIP-seq signal in GIST-T1 cells infected with Cas9 and either a control sgRNA directed at GFP (black, top) or sgRNAs directed against the two bound CTCF motifs in the KIT insulator (second row, red).

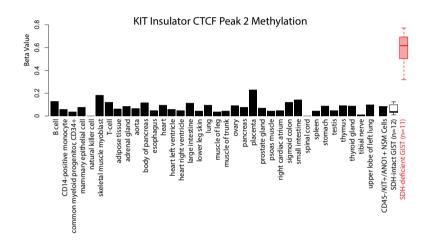


**Extended Data Fig. 6** | **Hypoxia marker induction in GIST-T1 cells. a**, Bar plot depicts levels of the TET product, 5-hydroxymethyl-cytosine (5-hmC), measured by ELISA in GIST-T1 cells infected with CRISPR-Cas9 and either a short guide RNA targeting GFP (sgGFP) or SDHB. Cells were cultured in either control media or media supplemented with 20  $\mu$ M of dimethylsuccinate (DMS), a membrane-permeable ester of succinate, for 3 days, as indicated. **b, c**, Plots show relative expression of pseudo-hypoxia-associated genes *EPAS1* (also known as *HIF2A*) $^{67}$ 

and  $IGF1R^{68}$  (**b**), and KITLG (also known as SCF) (**c**) in control GIST-T1 cells (black), SDH-deficient GIST-T1 cells generated by CRISPR-Cas9 knockout of SDHB and cultured with exogenous succinate (red), or GIST-T1 cells treated with the iron chelator DFX to simulate hypoxia (blue). Upregulation of KIT ligand due to pseudo-hypoxia or tumour hypoxia may supplement FGF ligands in promoting RTK signalling in SDH-deficient GIST.





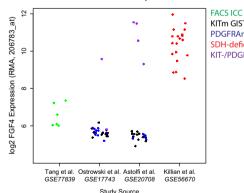


b

#### Mouse ICC Transcriptomic Data (FPKM)

Gene	Jejunal ICC	Colonic ICC
Fgf3	0.07	0.07
Fgf4	0.07	0
Ano1	702.72	1068.03
Kit	468.03	792
Kitl	13.48	57.33
Fgfr1	61.48	214

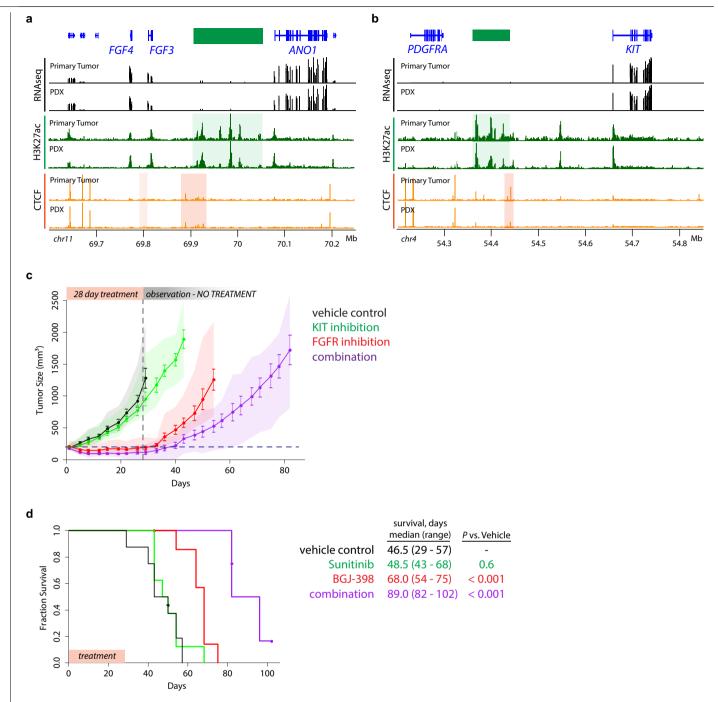
#### Human ICC/GIST U133 Microarray Data



KITM GIST
PDGFRAM GIST
SDH-deficient GIST
KIT-/PDGFRA-wildtype GIST (SDH status unknown)

 $\label{lem:continuous} \textbf{Extended Data Fig. 7} | \textbf{FGF and KIT insulator methylation and expression in} \\ \textbf{GIST subtypes and non-malignant cells. a}, \texttt{Bar plot depicts methylation of FGF} \\ \text{insulator CTCF peak 2 (top) and KIT insulator CTCF peak 2 (bottom) in 34 tissues and primary cells available through \texttt{ENCODE}^{60}. Values are average methylation of CpGs nearest the CTCF motifs, determined by whole genome bisulfite sequencing (WGBS) (KIT insulator, 3 CpGs; FGF insulator, 6 CpGs). Methylation of these sites is also shown for SDH-intact and SDH-deficient GISTs (see Fig. 3e, f), and for flow-sorted CD45^ANO1^KIT^* (ICC enriched) cells from normal stomach muscle (NSM) tissue ($n$ values represent biologically independent $n$ and $n$ are supposed to the continuous content of the conte$ 

specimens). **b**, Left, table depicts FPKM (fragments per kilobase of transcript per million mapped reads) values of relevant genes in mouse ICCs isolated from jejunum or colon<sup>61</sup>. Right, dot plot depicts expression of FGF4 in either flow sorted ICCs (green) or GISTs of the indicated subtype: *KIT* mutant in black, *PDGFRA* mutant in blue, SDH-deficient in red, and KIT-/PDGFRA wild type in purple<sup>62-66</sup>. SDH status of the latter group is unknown, but SDH-deficient GIST represent a significant portion of KIT-/PDGFRA wild-type tumours. Data are drawn from the indicated GEO series and publications.



**Extended Data Fig. 8 | PDX trial of FGFR and KIT combination therapy in SDH-deficient GIST. a, b,** Genomic views depict RNA expression (black), H3K27ac (green) and CTCF occupancy (orange) over the FGF (a) and KIT (b) loci for the S1 primary tumour and PDX. Genes (blue), super-enhancers (green bar and shade) and lost CTCF insulators (orange shade) are indicated. **c,** Plot depicts tumour volume during treatment and observation periods of experiment, as in Fig. 4g, except with time axis extended until final group reached censor point

(one tumour in the group >2,000 mm³). Points represent mean tumour size, error bars represent s.e.m., and shading represents range of tumour sizes for n=8 biologically independent xenograft-bearing mice per group. For statistics, see Fig. 4g. **d**, Kaplan–Meier plot depicts survival until clinical endpoint (tumour size >2,000 mm³) for the same PDX trial. Median and range survival are indicated for each group. P values reflect difference in survival between groups as calculated by logrank test.



Corresponding author(s):	Bradley Bernstein
Last updated by author(s):	Jul 27, 2019

## **Reporting Summary**

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, seeAuthors & Referees and theEditorial Policy Checklist.

$\overline{}$					
┖.	tっ	. +- :	İς	Ηт.	$\sim$
``	1 1			ш	· >

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.
n/a Confirmed
The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
🔲 🗷 A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
The statistical test(s) used AND whether they are one- or two-sided Only common tests should be described solely by name; describe more complex techniques in the Methods section.
A description of all covariates tested
🔲 🗷 A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i> ) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted Give <i>P</i> values as exact values whenever suitable.
For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i> ), indicating how they were calculated
Our web collection on <u>statistics for biologists</u> contains articles on many of the points above.
Software and code
Policy information about <u>availability of computer code</u>

Data analysis

Software usage and parameters are detailed in methods section of manuscript. Briefly, sequencing reads were aligned with BWA v. 0.7.4, Methylation data analyzed with methylCtools v. 0.9.4. ChIP-seq peaks were called and analyzed via Homer v. 4.9, MEME v. 4.7, HTSeq v. 0.6.152 and DESeq2 v. 1.16.1, general data analysis and graphing was performed in Matlab v. 9.1.0.441655, R v. 3.5.3, and IGV v. 2.5.3.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

#### Data

Policy information about availability of data

All manuscripts must include a <u>data availability statement</u>. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Sequencing data that support the findings of this study have been deposited in GEO with the accession code GSE107447.

Raw sequencing data generated through this project may contain identifiable human genetic information, as such it requires IRB approval to access (data has been deposited into a dbGaP dataset connected to the GEO database).

Raw data for the mouse xenograft trial (i.e. measured tumor volumes) are available as supplementary information table 3.

Field-spe	ecific reporting		
Please select the o	ne below that is the best fit for yo	ur research. If you are not sure, read the appropriate sections before making your selection.	
<b>x</b> Life sciences	Behavioural & social	I sciences	
For a reference copy of	the document with all sections, see <u>nature.c</u>	com/documents/nr-reporting-summary-flat.pdf	
Life scier	nces study desig	gn	
All studies must dis	sclose on these points even when	the disclosure is negative.	
Sample size	For analysis of clinical tissue, no statistical methods were used to predetermine sample size; rather, all available SDH-deficient tumor specimens with validated SDH loss and enough material available for analysis were tested. For mouse studies, no specific statistical calculations were performed; rather, sample size was determined based on prior experience with similar PDX trials.		
Data exclusions	No data were excluded from analyses.		
Replication	For all CRISPR insulator deletions, viral introduction of CRISPR/Cas9+sgRNA vector was repeated three times with separate viral preparations and infections to generate biologically independent replicates.  For epigenomic and transcriptomic characterization of clinical tissue (e.g. ChIP-seq, 4C, RNA-seq), multiple clinical specimens were analyzed, but technical replicates could not be performed on individual samples due to limited availability of material. All biological replicates (i.e. tumors of a given driver) were similar at reported sites (i.e. insulator loss/enhancer presence was consistent within driver subgroups).		
Randomization	For PDX trial, xenograft-bearing mice were randomized to treatment group at ~200mm3 tumor volume, with 8 mice per treatment group.		
Blinding	For PDX trial, blinding was not possible due to preparation and delivery methods of tested drugs. This is thought to have minimal impact on the studies, as no subjective (e.g. behavioral) criteria were measured.		
•	<u> </u>	aterials, systems and methods materials, experimental systems and methods used in many studies. Here, indicate whether each material,	
	· · · · · · · · · · · · · · · · · · ·	e not sure if a list item applies to your research, read the appropriate section before selecting a response.	
Materials & ex	perimental systems	Methods	
n/a Involved in the study		n/a Involved in the study	
Antibodies		ChiP-seq	
<b>x</b> Eukaryotic cell lines		Flow cytometry	
<b>▼</b> Palaeontology		MRI-based neuroimaging	

#### **Antibodies**

Antibodies used

Clinical data

Animals and other organismsHuman research participants

CTCF antibody is from Cell Signaling technologies, clone D31H2, catalog number 3418. Antibody was validated by manufacturer for ChIP in human cells, has been previously utilized by the authors (Flavahan et al., Nature 2016), and was validated as part of the ENCODE project via Western Blot for CTCF, as well as motif analysis to confirm enrichment for the known CTCF motif.

H3K27ac is a rabbit polyclonal antibody available from Active Motif, catalog number 39133, lot 31814008. Antibody was validated by the manufacturer for ChIP in human cells, and was validated as part of the ENCODE project.

Validation

Both antibodies were validated by manufacturers (including statements on the websites), and by the investigators and colleagues as part of the ENCODE project - including, but not limited to, western blot, immunoprecipitation, and ChIP motif finding and analysis of control cell lines and known peaks/motifs.

#### Eukaryotic cell lines

Policy information about cell lines

Cell line source(s) One cell line, GIST-T1, was used, and was obtained from the commercial vendor Cosmo

Biosciences

The cell line was not authenticated via STR testing, however locus sequencing for the KIT Authentication

gene confirmed the presence of the known and published KIT mutation present in GIST-T1

cells and the parental tumor the cell line was derived from.

The cell line was tested for mycoplasma via PCR-based method and confirmed to be Mycoplasma contamination

mycoplasma-free.

Commonly misidentified lines (See ICLAC register)

No commonly misidentified cell lines were utilized.

#### Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals For the PDX trial, female 8-week old NSG mice were utilized.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Wild animals No wild animals were utilized in this study.

No field-collected samples were utilized in this study. Field-collected samples

All animal experiments and study protocols were approved by the Dana Farber Cancer Institute Institutional Animal Care and Ethics oversight

Usage Committee (IACUC), and this is noted in the manuscript.

#### Human research participants

Policy information about studies involving human research participants

Population characteristics

The study involved the collection of deidentified and anonymized tumor material from patients at either Brigham and Women's Hospital, The Dana Farber Cancer Institute, or Massachusetts General Hospital. As such, no information about the patients, other than disease pathology, is known.

Patient tissue was obtained from tissue banks at either MGH or DFCI. Standard of care for GIST includes surgical resection of the tumor bulk - following treatment, excess surgical material from consenting patients was deposited into these banks. As the collection of these tissues occurs as part of the normal disease treatment, it is unlikely there is significant self-selection bias.

Ethics oversight The study protocol was approved by the Massachusetts General Hospital IRB and the Dana Farber Cancer Institute IRB.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

#### ChIP-seq

Recruitment

#### Data deposition

Confirm that both raw and final processed data have been deposited in a public database such as GEO.

Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links

May remain private before publication.

GEO database accession GSE107447

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE107447

Files in database submission

See included file, ChIP-seq\_2017-12-16052D\_file\_info.xlsx

Genome browser session (e.g. UCSC)

no longer applicable

Methodology

Replicates

For epigenomic and transcriptomic characterization of clinical tissue (e.g. ChIP-seq, 4C, RNA-seq), multiple clinical specimens were analyzed, but technical replicates could not be performed on individual samples due to limited availability of material. All clinical samples within a driver subgroup (SDH-intact vs. SDH-deficient) were highly similar at tested locations (e.g. KIT/ FGF insulator loss and superenhancer presence). See extended data figure 2 for more info.

Sequencing depth

See included file, ChIP-seq 2017-12-16052D file info.xlsx, which includes sequencing depth for each experiment.

Antibodies

CTCF antibody is from Cell Signaling technologies, clone D31H2, catalog number 3418. Antibody was validated by manufacturer for ChIP in human cells, has been previously utilized by the authors (Flavahan et al., Nature 2016), and was validated as part of the ENCODE project via Western Blot for CTCF, as well as motif analysis to confirm enrichment for the known CTCF motif.

H3K27ac is a rabbit polyclonal antibody available from Active Motif, catalog number 39133, lot 31814008. Antibody was validated by the manufacturer for ChIP in human cells, and was validated as part of the ENCODE project.

Peak calling parameters

Peaks were called with HOMER 4.9 against input controls. To call all H3K27ac peaks, we used 'histone' settings. To call superenhancers, we used 'super' settings and no local filtering. CTCF peaks were called with 'factor' settings.

Data quality

All reported peaks are detected with FDR < 0.1% and fold change > 4

Software

Homer 4.9, DESeq2 1.16.1, FIMO/MEME 4.7, HTSeq 0.6.1, featureCounts 1.6.2, and CNVnator 0.3

# Structure of the Fanconi anaemia monoubiquitin ligase complex

https://doi.org/10.1038/s41586-019-1703-4

Received: 26 April 2019

Accepted: 18 September 2019

Published online: 30 October 2019

Shabih Shakeel<sup>1,5</sup>, Eeson Rajendra<sup>1,5</sup>, Pablo Alcón<sup>1</sup>, Francis O'Reilly<sup>2</sup>, Dror S. Chorev<sup>3</sup>, Sarah Maslen<sup>1</sup>, Gianluca Degliesposti<sup>1</sup>, Christopher J. Russo<sup>1</sup>, Shaoda He<sup>1</sup>, Chris H. Hill<sup>1</sup>, J. Mark Skehel<sup>1</sup>, Sjors H. W. Scheres<sup>1</sup>, Ketan J. Patel<sup>1</sup>, Juri Rappsilber<sup>2,4</sup>, Carol V. Robinson<sup>3</sup> & Lori A. Passmore<sup>1\*</sup>

The Fanconi anaemia (FA) pathway repairs DNA damage caused by endogenous and chemotherapy-induced DNA crosslinks, and responds to replication stress<sup>1,2</sup>. Genetic inactivation of this pathway by mutation of genes encoding FA complementation group (FANC) proteins impairs development, prevents blood production and promotes cancer<sup>1,3</sup>. The key molecular step in the FA pathway is the monoubiquitination of a pseudosymmetric heterodimer of FANCD2-FANCI<sup>4,5</sup> by the FA core complex—a megadalton multiprotein E3 ubiquitin ligase<sup>6,7</sup>. Monoubiquitinated FANCD2 then recruits additional protein factors to remove the DNA crosslink or to stabilize the stalled replication fork. A molecular structure of the FA core complex would explain how it acts to maintain genome stability. Here we reconstituted an active, recombinant FA core complex, and used cryo-electron microscopy and mass spectrometry to determine its structure. The FA core complex comprises two central dimers of the FANCB and FA-associated protein of 100 kDa (FAAP100) subunits, flanked by two copies of the RING finger subunit, FANCL. These two heterotrimers act as a scaffold to assemble the remaining five subunits, resulting in an extended asymmetric structure. Destabilization of the scaffold would disrupt the entire complex, resulting in a non-functional FA pathway. Thus, the structure provides a mechanistic basis for the low numbers of patients with mutations in FANCB, FANCL and FAAP100. Despite a lack of sequence homology, FANCB and FAAP100 adopt similar structures. The two FANCL subunits are in different conformations at opposite ends of the complex, suggesting that each FANCL has a distinct role. This structural and functional asymmetry of dimeric RING finger domains may be a general feature of E3 ligases. The cryo-electron microscopy structure of the FA core complex provides a foundation for a detailed understanding of its E3 ubiquitin ligase activity and DNA interstrand crosslink repair.

The FA core complex is composed of eight stably-associated subunits: FANCA, FANCB, FANCC, FANCE, FANCF, FANCG, FANCL and FAAP100<sup>6</sup>. FANCL contains a RING-finger domain which acts as the E3 ubiquitin ligase. It associates with FANCB and FAAP100 to form a catalytic module<sup>6,8</sup>; a low-resolution negative-stain electron microscopy (EM) study suggested that, in the absence of the other subunits, this is a symmetric dimer of FANCB–FANCL–FAAP100 heterotrimers<sup>9</sup>. FANCA and FANCG are proposed to act as a chromatin-targeting module, whereas FANCC, FANCE and FANCF form a substrate-recognition module<sup>6,8,10,11</sup>. Despite the central role of the FA core complex in DNA repair, we lack a molecular understanding of how FANCL incorporates into the complex to perform site-specific monoubiquitination of the FANCD2–FANCI substrate and how mutation disrupts the function of the complex<sup>12</sup>.

To determine the structure of the FA core complex, we overexpressed all eight subunits from  $Gallus\ gallus\ (chicken)$  on a single baculovirus

in insect cells, which enabled us to purify an intact, recombinant complex (Fig. 1a). The purified complex specifically monoubiquitinated FANCD2 but not FANCI in vitro (Extended Data Fig. 1), similar to the native chicken complex  $^6$ .

To investigate the molecular basis of subunit association, we imaged this recombinant FA core complex using cryo-EM. This revealed an elongated structure, about 25 nm in length (Fig. 1b, Extended Data Fig. 2a). We determined a 3D reconstruction of the FA core complex at an overall resolution of 4.2 Å (Extended Data Fig. 2b–e, Extended Data Table 1). The peripheral regions were less well resolved than the central core, possibly owing to conformational flexibility. In agreement with conformational heterogeneity in the complex, we detected structural variations using multi-body refinement<sup>13</sup>, including a continuum of conformational movement of the top and base regions (Extended Data Fig. 2f, Supplementary

<sup>1</sup>MRC Laboratory of Molecular Biology, Cambridge, UK. <sup>2</sup>Bioanalytics, Institute of Biotechnology, Technische Universität Berlin, Berlin, Germany. <sup>3</sup>Physical and Theoretical Chemistry Laboratory, University of Oxford, Oxford, UK. <sup>4</sup>Wellcome Centre for Cell Biology, University of Edinburgh, Edinburgh, UK. <sup>5</sup>These authors contributed equally: Shabih Shakeel, Eeson Rajendra. \*e-mail: passmore@mrc-lmb.cam.ac.uk

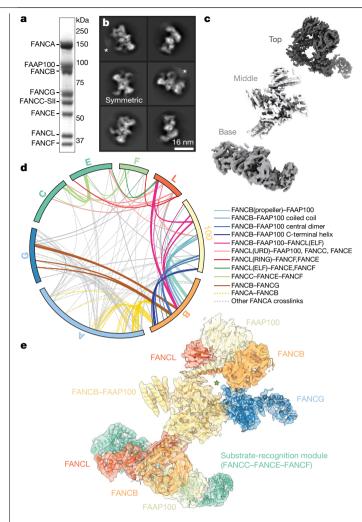


Fig. 1 | Overall structure of the FA core complex. a, SDS-PAGE analysis of purified FA core complex with subunits and molecular weight markers indicated. FANCC carries a 2× Strep II tag on its C terminus (FANCC-SII). This purification was repeated more than three times with similar results. For gel source data, see Supplementary Fig. 1. b, Selected 2D reference-free class averages of the FA core complex. One class appears to be symmetric (labelled). Asterisks mark disordered density extending from the side of the complex that does not align well. c, Focused classification and refinement on the top and base regions, and multibody refinement on the middle region resulted in three independent cryo-EM maps that are shown separately, in three different shades of grey. d, Crosslinking mass spectrometry revealed 834 crosslinks (1% false discovery rate) between residues that are in close proximity. Intermolecular crosslinks are shown, coloured by interacting regions. e, Model of FA core complex (cartoon representation) fitted into the EM density (isosurface representation with transparency). Map and model are coloured by assigned subunits. The green star marks a channel with diameter approximately 23  $\hbox{\AA}.$ 

Videos 1, 2). Particle subtraction followed by focused classification and refinement<sup>14</sup> generated separate, improved reconstructions of the top and base regions (Fig. 1c, Extended Data Fig. 2g, h and Supplementary Video 3).

Our map of the complete FA core complex was of sufficient resolution to dock existing structures and to resolve secondary structure elements (Extended Data Fig. 2i). We fit two previously determined high-resolution structures (FANCL<sup>15</sup> and part of FANCF<sup>16</sup>) into the map, accounting for about 12% of the entire mass of the complex. Most subunits do not have substantial homology to proteins of known structure (Extended Data Table 2), so we modelled  $\alpha$ -helices and  $\beta$ -strands into the remainder of the map. To determine which subunits these  $\alpha$ -helices and  $\beta$ -strands belong to, we required additional data.

Next, we purified subcomplexes and imaged them using cryo-EM. By comparing the 2D class averages of subcomplexes to the complete FA core complex, we identified regions corresponding to specific subunits (Extended Data Fig. 3a-c). Removal of FANCA did not substantially change the class averages. This suggested that FANCA may be conformationally heterogeneous and blurred out in reconstructions, or it may dissociate or denature during cryo-EM specimen preparation. The base was absent in a complex of FANCA, FANCG, FANCB, FANCL and FAAP100, suggesting that the base probably contains the substrate-recognition module (FANCC, FANCE and FANCF). The partially disordered arm that extends from the central part of the complex is probably FANCG because this density is lost when FANCG is removed from a FANCB-FANCL-FAAP100 complex. Finally, 2D classes of the catalytic module (FANCB, FANCL and FAAP100) resemble the middle region of the FA core complex.

We also studied the structure of the FA core complex using noncovalent native mass spectrometry (Extended Data Fig. 3d, e). During ionization, the FANCE subunit tends to dissociate and subcomplexes are formed, providing information on subunit stoichiometry and proteinprotein interactions. The largest complex (808 kDa) that we detected in native mass spectrometry contains seven of the eight different subunits, including two copies of FANCB, FANCL and FAAP100, and a single copy of each of the remaining subunits. This agrees with the subunit stoichiometries of a purified native FA core complex<sup>6</sup>. FANCB, FANCL and FAAP100 were present in most of the subcomplexes identified by native mass spectrometry (Extended Data Fig. 3d), suggesting that they form a central core.

To identify residues in close proximity, we performed crosslinking mass spectrometry (Fig. 1d, Extended Data Fig. 4). This revealed 834 inter- and intramolecular crosslinks, with 40% of these located in FANCB. FANCL and FAAP100. This is consistent with these three subunits forming an intimate complex. By combining the crosslinking mass spectrometry data showing which residues are in close proximity with the subunit assignment from subcomplexes, the subunit stoichiometry from native mass spectrometry, homology modelling and secondary structure predictions, we generated models for all FA core complex subunits except FANCA (Fig. 1e, Extended Data Fig. 5a-d, Extended Data Table 1, 2, Supplementary Video 4, Methods).

A dimer of FANCB-FAAP100 heterodimers is located in the middle region of the structure (Fig. 2). Two pairs of long  $\alpha$ -helices (coiled coils) connect central \( \beta \)-strands and helical bundles with peripheral densities (Fig. 2b. c), Crosslinking mass spectrometry and modelling showed that each coiled coil is probably composed of α-helices from FANCB and FAAP100 (Extended Data Fig. 5e, f). At the peripheral ends of the coiled coils, we identified pairs of  $\beta$ -propellers, each containing a  $\beta$ -propeller from the N-terminal region of FANCB or FAAP100 (Fig. 2d). We could  $differentiate \ the \ two \ \beta\text{-propellers} \ on \ the \ basis \ of \ cross links: the \ FANCB$ β-propeller is near the coiled coil, whereas the FAAP100 β-propeller is close to the ELF domain of FANCL. Unexpectedly, despite the lack of sequence homology, FANCB and FAAP100 share markedly similar overall structures and domain organizations (Fig. 2e).

A homology model of FANCL, including the ELF, URD and RING finger domains<sup>15</sup>, fits into the base of the complex (Fig. 3a, b) but the relative orientations of the individual domains are different compared with the crystal structure (Extended Data Fig. 4b). By contrast, only the ELF domain could be placed into the second copy of FANCL at the top of the complex (Fig. 3c). Hydrogen-deuterium exchange mass spectrometry (HDX-MS) confirmed that FANCL interacts with the FANCB-FAAP100 coiled coil (Extended Data Fig. 6).

FANCG contains tetratricopeptide repeats (TPRs) (Fig. 3d, e),  $cross links to the central FANCB-FAAP 100\,dimer\,(Fig. 1d), and is required$ for FANCA association with the FA core complex (Extended Data Fig. 3b). Of note, there is a channel between FANCG and the catalytic module (Fig. 1e). FANCA, which is absent in the maps, is probably peripheral to FANCG, and possibly located in the blurred density visible in 2D class averages (indicated by asterisks in Fig. 1b; Extended Data Fig. 5d).

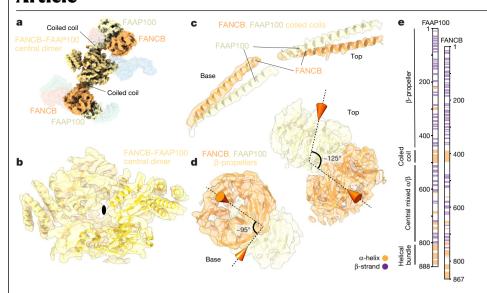


Fig. 2 | The molecular scaffold of the FA core complex includes a dimer of FANCB-FAAP100 heterodimers, a. Surface representation of the FA core complex model, highlighting FANCB and FAAP100. FANCB, orange; FAAP100, yellow; regions where we are unable to distinguish FANCB and FAAP100, yellow-orange. b-d. Models of FANCB and FAAP100 subunits in cartoon representation placed into the cryo-EM map. In b, a black oval marks the pseudo two-fold symmetry axis. There are substantial differences between the two symmetry-related copies, which are shown in different shades of vellow in the model. In d, cones indicate the orientations of the central pores of the β-propellers and the angles between them are shown. e, Proposed domain organization of FANCB and FAAP100, showing their structural similarity.

The large number of crosslinks between FANCA and FANCG are consistent with their proximity (Fig. 1d).

The substrate-recognition module (FANCC, FANCE and FANCF), located in the base, comprises an arc of  $\alpha$ -helices (Fig. 3f). FANCF occupies a central position within the arc. Crosslinking mass spectrometry showed that the FANCL RING finger and ELF domain contact FANCE and FANCF, the FANCL URD domain contacts FANCC and FANCE, and the FANCC C-terminal region and FANCE are in close proximity. Native mass spectrometry also showed direct contact between FANCE and FANCF (Extended Data Fig. 3d).

We also determined a structure of a subcomplex, present at a lower abundance in our sample, at an overall resolution of 4.6 Å (Extended Data Fig. 7a–d). This subcomplex was symmetric but the map did not improve on application of C2 symmetry. We therefore implemented a local symmetry algorithm in Relion for averaging the two halves of the subcomplex map (Methods, Extended Data Fig. 7e, f). This symmetric structure revealed an assembly comprising two copies of each of FANCB, FANCL, FAAP100 and FANCG (Extended Data Fig. 8a–c, Supplementary Video 5). It is unclear whether this subcomplex has a functional role in vivo or whether it is an assembly intermediate. In both copies

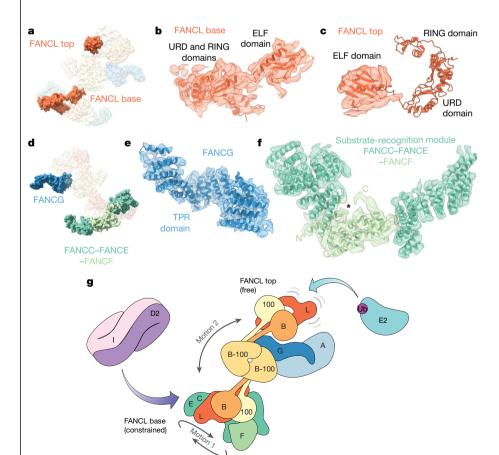


Fig. 3 | Asymmetric dimerization in the FA core complex.a-c, Models of FANCL in the FA core complex. a, Surface representation of the FA core complex model, highlighting the two copies of  $\mathsf{FANCL}.\,\bm{b},\bm{c},\mathsf{Models}\,\mathsf{of}\,\mathsf{FANCL}_\mathsf{base}\,\mathsf{and}\,\mathsf{FANCL}_\mathsf{top}$ subunits in cartoon representation are shown fitted in the cryo-EM map. Density for the URD and RING domains is not well defined in the top copy (c).  $\textbf{d-f}, Models \, of FANCG \, and \, the substrate-recognition$ module. d, Surface representation of the FA core complex model highlighting FANCG and the substrate-recognition module (FANCC-FANCE-FANCF). e, f, Models of FANCG TPR domain (e) and the substrate-recognition module (  $\mathbf{f}$  ) are shown fitted into the cryo-EM map. The crystal structure of FANCF could be assigned and the N and C termini of the model are indicated. The first helix (N to \*) was not present in the crystal structure. Since all three subunits of the substrate-recognition module are substantially helical, it was not possible to assign the remaining helices of the base to individual subunits. g, Model for monoubiquitination of FANCD2 by the FA core complex. The major motions detected in multibody refinement are indicated with grey arrows.

of FANCL, the URD and RING finger domains have weak density or are not visible, similar to the top FANCL (FANCL<sub>top</sub>) in the full, asymmetric FA core complex.

Comparison of subcomplex and FA core complex structures suggests that the substrate-recognition module alters the relative orientations of the β-propellers and coiled coil in the base (Fig. 2d, Supplementary Video 6). This disrupts the symmetry of the catalytic module in the complete FA core complex, provides a binding site in the base for the URD and RING finger domains of FANCL and probably disrupts docking of a second (symmetric) copy of FANCG onto the middle region owing to steric clashes (Extended Data Fig. 8d). These structural alterations may be transmitted to the top region to prevent the binding of a second substrate-recognition module, consistent with allosteric coupling proposed previously<sup>9</sup>. In agreement with this, a purified substrate-recognition module did not readily associate with the symmetric subcomplex in vitro to form the asymmetric FA core complex (Extended Data Fig. 8e, f), suggesting that in vivo assembly is required.

Dimerization is required for the activity of other E3 ligases, including the Rad18, RNF8 and CHIP homodimers<sup>17-19</sup> and the BRCA1-BARD1 and Ring1b-Bmi1 heterodimers<sup>20,21</sup>. These contain two RING finger-U-box domains arranged in an asymmetric manner with only a single functional E2 binding site. Notably, the activity of multi-subunit cullin-RING finger E3 ligases is stimulated by dimerization<sup>22,23</sup>. Thus, structural and functional asymmetry appear to be a common feature of E3 ligases but the spatial separation of the FANCL RING fingers in the FA core complex is unusual (Extended Data Fig. 9a, b). Like other dimeric RING finger E3s, one RING finger (in the FANCL subunit at the base (FANCL<sub>base</sub>)) of the FA core complex may have a structural role in promoting substrate binding along with FANCE<sup>10,11,24,25</sup>, whereas the other (FANCL<sub>top</sub>) may be the active E3 that promotes ubiquitin transfer<sup>26</sup> to FANCD2 (Fig. 3g) within the structural state we observe. The FANCD2-FANCI substrate is also an asymmetric dimer. Thus, each of the two FANCL RING finger domains could monoubiquitinate one of the substrate proteins. Nevertheless, this purified complex does not monoubiquitinate FANCI, so an additional activation step might be required. Substrate binding is not required to activate the E3 ligase activity (Extended Data Fig. 9c, d).

The majority of FA-complex mutations detected in patients with FA result in protein truncation and are found in the structural periphery of the FA core complex<sup>27</sup> (Extended Data Fig. 9e). Residual monoubiquitin ligase activity is still present after deletion of peripheral subunits in cells<sup>6,8</sup> and in vitro (Extended Data Fig. 9f), indicating a partially functioning core complex. By contrast, deletion of FANCB, FANCL or FAAP100, which comprise the catalytic module and the structural scaffold for the complex, eliminates this residual activity<sup>6,8</sup>. Patients with FA who carry mutations in FANCB or FANCL are severely afflicted, and we predict that FANCB missense mutations (L43S, P230S, L329P and L676P) disrupt stability of the catalytic module (indicated by asterisks in Extended Data Fig. 4a). Together, these data provide genetic and clinical support that mutations in the catalytic module result in disruption of the FA core complex structure. By contrast, mutations in the periphery do not prevent complex formation and may be better tolerated.

In summary, our data provide a structural model for the FA core complex, enabling an interpretation of the molecular pathophysiology of FA. The reconstituted system we describe will also enable further mechanistic questions to be addressed, including precisely how this large complex functions as a DNA damage-inducible monoubiquitin ligase.

### Online content

Any methods, additional references. Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1703-4.

- Crossan, G. P. & Patel, K. J. The Fanconi anaemia pathway orchestrates incisions at sites of crosslinked DNA. J. Pathol. 226, 326-337 (2012).
- Schlacher, K., Wu, H. & Jasin, M. A distinct replication fork protection pathway connects Fanconi anemia tumor suppressors to RAD51-BRCA1/2. Cancer Cell 22, 106-116 (2012).
- Nalepa, G. & Clapp, D. W. Fanconi anaemia and cancer: an intricate relationship. Nat. Rev. Cancer 18, 168-185 (2018).
- Knipscheer, P. et al. The Fanconi anemia pathway promotes replication-dependent DNA interstrand cross-link repair. Science 326, 1698-1701 (2009).
- Smogorzewska, A. et al. Identification of the FANCI protein, a monoubiquitinated FANCD2 paralog required for DNA repair. Cell 129, 289-301 (2007).
- Raiendra, E. et al. The genetic and biochemical basis of FANCD2 monoubiquitination. Mol. Cell 54, 858-869 (2014).
- Garcia-Higuera, Let al. Interaction of the Fanconi anemia proteins and BRCA1 in a common pathway, Mol. Cell 7, 249-262 (2001).
- Huang, Y et al. Modularized functions of the Fanconi anemia core complex. Cell Rep. 7. 1849-1857 (2014)
- Swuec, P. et al. The FA core complex contains a homo-dimeric catalytic module for the symmetric mono-ubiquitination of FANCI-FANCD2, Cell Rep. 18, 611-623 (2017).
- 10. Pace, P. et al. FANCE: the link between Fanconi anaemia complex assembly and activity. EMBO J. 21, 3414-3423 (2002).
- 11. van Twest, S. et al. Mechanism of ubiquitination and deubiquitination in the Fanconi anemia pathway. Mol. Cell 65, 247-259 (2017).
- Walden, H. & Deans, A. J. The Fanconi anemia DNA repair pathway: structural and functional insights into a complex disorder. Annu. Rev. Biophys. 43, 257-278 (2014).
- Nakane, T., Kimanius, D., Lindahl, E. & Scheres, S. H. Characterisation of molecular motions in cryo-EM single-particle data by multi-body refinement in RELION. eLife 7,
- Bai, X. C., Rajendra, E., Yang, G., Shi, Y. & Scheres, S. H. Sampling the conformational space of the catalytic subunit of human y-secretase. eLife 4, e11182 (2015).
- Cole, A. R., Lewis, L. P. C. & Walden, H. The structure of the catalytic subunit FANCL of the Fanconi anemia core complex. Nat. Struct. Mol. Biol. 17, 294-298 (2010).
- Kowal, P., Gurtan, A. M., Stuckert, P., D'Andrea, A. D. & Ellenberger, T. Structural determinants of human FANCF protein that function in the assembly of a DNA damage signaling complex. J. Biol. Chem. 282, 2047-2055 (2007).
- Huang, A. et al. Symmetry and asymmetry of the RING-RING dimer of Rad18. J. Mol. Biol. 410, 424-435 (2011).
- Mattiroli, F. et al. RNF168 ubiquitinates K13-15 on H2A/H2AX to drive DNA damage signaling, Cell 150, 1182-1195 (2012).
- Zhang, M. et al. Chaperoned ubiquitylation—crystal structures of the CHIP U box E3 ubiquitin ligase and a CHIP-Ubc13-Uev1a complex. Mol. Cell 20, 525-538 (2005).
- Brzovic, P. S., Rajagopal, P., Hoyt, D. W., King, M. C. & Klevit, R. E. Structure of a BRCA1-BARD1 heterodimeric RING-RING complex. Nat. Struct. Biol. 8, 833-837 (2001)
- Buchwald, G. et al. Structure and E3-ligase activity of the Ring-Ring complex of polycomb proteins Bmi1 and Ring1b. EMBO J. 25, 2465-2474 (2006)
- Passmore, L. A. et al. Structural analysis of the anaphase-promoting complex reveals multiple active sites and insights into polyubiquitylation. Mol. Cell 20, 855-866 (2005)
- Tang, X. et al. Suprafacial orientation of the SCFCdc4 dimer accommodates multiple geometries for substrate ubiquitination. Cell 129, 1165-1176 (2007).
- Gordon, S. M., Alon, N. & Buchwald, M. FANCC, FANCE, and FANCD2 form a ternary complex essential to the integrity of the Fanconi anemia DNA damage response pathway. J. Biol. Chem. 280, 36118-36125 (2005).
- Polito, D. et al. The carboxyl terminus of FANCE recruits FANCD2 to the Fanconi Anemia (FA) E3 ligase complex to promote the FA DNA repair pathway. J. Biol. Chem. 289,
- Zimmerman, E. S., Schulman, B. A. & Zheng, N. Structural assembly of cullin-RING ubiquitin ligase complexes. Curr. Opin. Struct. Biol. 20, 714-721 (2010).
- Neveling, K., Endt, D., Hoehn, H. & Schindler, D. Genotype-phenotype correlations in Fanconi anemia. Mutat. Res. 668, 73-91 (2009).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

# Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

### Cloning, expression and purification

cDNAs encoding full length G. gallus (Gg) FANCA, FANCB, FANCC, FANCE, FANCF, FANCG, FANCL and FAAP100 were synthesized (GeneArt). FANCC contained a C-terminal extension with a 3C protease site and double Strep II tag. For protein expression, all genes were cloned into the MultiBac expression system and constructs were generated as previously described<sup>28</sup>. In brief, to make gene expression cassettes, FANCA and FANCG were subcloned via BamHI and XbaI into pACEBac1: FANCE, FANCE and FAAP100 were subcloned via Xhol and Kpnl into pIDS; and FANCB and FANCL were subcloned via BamHI and XbaI into pIDC. Gene cassettes were then sequentially subcloned as BstXI-I-Ceul or BstXI-PI-Scel fragments into I-Ceul or PI-Scel sites to generate pACEBac1-FANCA-FANCG, pIDS-FANCC-3C-2×StrepII-FANCE-FANCF, and pIDS-FAAP100-FANCB-FANCL. The spectinomycin antibiotic resistance cassette of pIDS-FAAP100-FANCB-FANCL was substituted with the kanamycin antibiotic resistance cassette of pIDK as a SnaBI-PI-SceI fragment to generate pIDK-FAAP100-FANCB-FANCL.

pACEBac1-FANCA-FANCG, pIDK-FAAP100-FANCB-FANCL and pIDS-FANCC-3C-2×StrepII-FANCE-FANCF were fused using Cre recombinase (NEB) to generate a single vector (gentamicin, spectinomycin and kanamycin resistant) containing a single copy of each gene (FA core complex) used to make protein for the initial data collection. All constructs were confirmed by restriction digest analysis, PCR and sequencing.

For subsequent protein preparations, A-G-B-L-100-C-E-F, A-G-B-L-100, G-B-L-100 and C-E-F complexes (letters indicate the FANC family members) were prepared using a modified BiGBac system as described previously<sup>29,30</sup>. The individual genes were PCR amplified for cloning into pBIG vectors from pACEBac1, pIDC or pIDS vectors. Sequences encoding 2×Strep II tag and 3C protease site were included on *FANCC*. If *FANCC* was not present, the tag sequence was added to *FANCB*. The combined vector carrying the FA core complex or subcomplex was transformed into EMBacY cells to generate a bacmid. Bacmid DNA was transfected into Sf9 cells and virus was passaged twice in the same cell line before large-scale infection in Sf9 cells. Infected cells were collected when cell growth arrested. Sf9 cells were obtained from Oxford Expression Technologies, catalogue no. 600100 (negative for mycoplasma, identity not independently authenticated by us).

Cells were lysed by sonication in lysis buffer ( $100\,\text{mM}$  HEPES pH 8.0,  $300\,\text{mM}$  NaCl,  $1\,\text{mM}$  TCEP, 5% glycerol, EDTA-free protease inhibitor,  $5\,\text{mM}$  benzamidine hydrochloride and  $100\,\text{U}\,\text{ml}^{-1}$  benzonase). Clarified cell lysate was incubated with StrepTactin resin (GE Healthcare) for  $1\,\text{h}$  followed by wash with lysis buffer. Proteins were eluted in elution buffer ( $100\,\text{mM}$  HEPES pH 8.0,  $300\,\text{mM}$  NaCl,  $1\,\text{mM}$  TCEP, 5% glycerol and  $8\,\text{mM}$  desthiobiotin). Further purification was performed by HiTrap Heparin HP affinity column (GE Healthcare) using a linear gradient of NaCl from concentration of  $150\,\text{mM}$  to  $1\,\text{M}$  in  $50\,\text{mM}$  HEPES pH 8.0,  $1\,\text{mM}$  TCEP, over  $22\,\text{column}$  volumes. For FA core complex, this was followed by anion-exchange chromatography (MonoQ, GE Healthcare) in the same buffer using a linear gradient of NaCl from concentration of  $150\,\text{mM}$  to  $1\,\text{M}$  over  $20\,\text{column}$  volumes. The final buffer for purified FA core complex and subcomplexes was  $50\,\text{mM}$  HEPES pH 8.0,  $-500\,\text{mM}$  NaCl,  $1\,\text{mM}$  TCEP.

# **Ubiquitination assay**

Ubiquitination assays were performed as described previously  $^{6,31}$ . In brief, a reaction volume of 20  $\mu$ l contained 75 nM E1 (Boston Biochem), 1  $\mu$ M E2 (GgUBE2T), 0.25  $\mu$ M E3 (FA core complex), 1  $\mu$ M substrate (His-GgFANCI, GgFANCI (KS2SR, His-GgFANCD2, GgFANCD2 (SG3R), 50  $\mu$ M 5′-flapped DNA and 20  $\mu$ M haemagglutinin (HA)–ubiquitin (Boston Biochem). For ubiquitin discharge assays, concentrations of 125 nM

E1 (Boston Biochem),  $5\,\mu\text{M}$  E2 (GgUBE2T),  $1.5\,\mu\text{M}$  E3 (FA core complex) and 50 mM free lysine (instead of substrate) were used. The  $0.25\,\mu\text{M}$  E3 enzyme concentration is based on the FANCL subunit, estimated by comparing the amount of FANCL in FA core complex and subcomplexes against purified FANCL of known concentration on SDS–PAGE. The reaction buffer was 50 mM HEPES pH 8.0, 64 mM NaCl, 4% glycerol, 5 mM MgCl<sub>2</sub>, 2 mM ATP and 0.5 mM DTT. The reaction was incubated at 30 °C for 90 min, stopped by adding NuPAGE LDS sample buffer (Thermo Fisher) and run on an SDS polyacrylamide gel (3–8% NuPAGE Tris-acetate). Samples were analysed by Coomassie staining or by western blot using a HA antibody (Santa Cruz Biotechnology). All assays were performed independently three times.

### **Electron microscopy**

Protein complexes were vitrified by applying 3–3.5  $\mu$ l purified protein (-1 $\mu$ M) to UltraAuFoil R1.2/1.3 grids (Quantifoil)<sup>32</sup> with a thin continuous carbon support layer (for initial FA core complex dataset) or in unsupported ice (for final FA core complex dataset and subcomplexes) that had been made hydrophilic using an argon:oxygen plasma, blotting for 2.5, 3 or 4.5 s at 4 °C with relative humidity of 100% and plunging into liquid ethane using a Vitrobot Mark IV (FEI).

Cryo-EM data were collected on a FEI Titan Krios transmission electron microscope operated at 300 keV acceleration voltage using EPU automated data collection software. An initial dataset was collected on a Falcon II detector at 47,000× nominal magnification with a pixel size of 1.774 Å per pixel. The final data for FA core complex (Extended Data Table 1) were collected on a Falcon III detector in counting mode at 75,000× nominal magnification, and pixel size of 1.04 Å (MRC LMB) or 1.085 Å (eBIC). The subcomplexes G-B-L-100-C-E-F, A-G-B-L-100 and G-B-L-100 were imaged at 59,000× nominal magnification on a Falcon III detector in integrating mode. B-L-100 data were collected at 47,000× on a Falcon II detector in integrating mode.

### **Cryo-EM image processing**

Image processing was performed in Relion (v.2 and v.3.0-beta)<sup>33-35</sup>, and Relion wrappers were used for external programs except for EMAN2<sup>36</sup>. For all datasets, whole-frame alignment was performed using Motion-Cor2<sup>37</sup>, and contrast transfer function parameters were estimated using gCTF<sup>38</sup>. 3D maps were post-processed to automatically estimate and apply the B-factor and to determine the resolution by Fourier shell correlation (FSC) between two independent half datasets using 0.143 criterion<sup>39</sup>. Local resolution was estimated using ResMap<sup>40</sup>.

### Initial model

The initial FA core complex dataset was processed in Relion v.2. Particles were picked manually from a few micrographs, and used for 2D class averaging with a box size of 390 pixels. The resulting 2D class averages were used to pick particles from all micrographs using template-matching in Relion's autopicker. After 2D classification of the auto-picked particles, selected classes were used to make an initial 3D model in EMAN2 $^{36}$ . This initial model was used for subsequent 3D classification and refinement in Relion.

### Refinement

The map generated from the Falcon II dataset was used during the first round of 3D classification with particles from the Falcon III dataset of FA core complex. The datasets from MRC LMB and eBIC were initially processed separately, to generate separate 3D reconstructions. The pixel size of eBIC data was determined using the 2.35 Å spacing from the gold foil images collected under the same imaging conditions as the sample data. The eBIC micrographs were then rescaled to the pixel size of LMB data and CTF was re-estimated, and all datasets were merged.

The FA core complex map was divided into three bodies: body  $1 \pmod{1}$  (middle region), body  $2 \pmod{2}$  (base region) and body  $3 \pmod{9}$  (top region). Bodies  $2 \pmod{3}$  were rotated relative to  $1 \pmod{9}$  (middle) region appears to be

the most rigid (Extended Data Fig. 2f). Sigma angles of 10 and a sigma offset of 2 were used during refinement. Multibody refinement<sup>13</sup> was continued from the last iteration of consensus refinement. The middle region was best resolved with an overall resolution of 4.4 Å whereas the top and base were ~7 Å resolution. To estimate the flexibility in the FA core complex we performed principal component analysis on the optimal orientations of all the bodies for all particle images in the dataset using relion\_flex\_analyse<sup>13</sup>. We rendered videos for principal components 1 and 2 as these described ~30% of the variance in the rotations and translations.

Per-particle CTF refinement and beam tilt estimation were performed by dividing the datasets into individual data collection sessions and processing in a bigger box of 586 pixels, followed by further refinement of the resolution of the maps for the top and base regions of the complete complex were further improved by performing focused classification with signal subtraction followed by refinement (Extended Data Fig. 2g, h). The overall resolution is probably limited by heterogeneity. A regularization T-value of 5 was used during 3D refinement to boost the contribution of higher spatial frequencies to improve the quality of the map to aid in model building.

The subcomplexes were processed up to 2D classification as there were not enough different views to generate a 3D map.

## Local symmetry averaging

Cryo-EM structure determination by single-particle analysis relies on the reduction of noise through averaging over multiple copies of extremely noisy projection images of individual macromolecular complexes. For symmetric complexes, for example, for homo-multimers or for icosahedral virus capsids, additional averaging can be performed by imposing point-group symmetry on the reconstruction. Because each projection image of a symmetrical object provides multiple views of the asymmetric unit, compared to asymmetric complexes, the same number of images will yield a better reconstruction, or fewer images are needed to obtain a reconstruction of the same quality. Therefore, point-group symmetry averaging is commonly employed in single-particle analysis refinement programs.

The FA subcomplex described here does not obey overall point-group symmetry, but still contains multiple, potentially identical, subcomplexes. We call this local symmetry. Averaging over locally symmetric subcomplexes in cryo-EM single-particle analysis has previously been performed to improve reconstructions after completion of the refinement process, for example on the subunits of triangulation number T>1 virus capsids 42.43. However, local symmetry averaging has the potential to improve the reconstruction at every stage of the iterative refinement process. Better reconstructions during refinement will lead to better alignments, and hence a better final reconstruction. Therefore, we implemented a local symmetry averaging approach inside the relion\_refine program. This approach has conceptual similarities to imposing noncrystallographic symmetry (NCS) in X-ray crystallography 44.

This new implementation within Relion 3.0 allows the user to define an arbitrary number of groups, each with an arbitrary number of assumed identical subcomplexes. For each group, the user provides a mask around one member of the group, as well as 3D transformation matrices (expressed as three Euler angles and three translations in x, y and z) in real space to superimpose that member onto each of the other members in the group. This information is expressed in a Relion-style STAR file, which is passed using the local\_symmetry command line option to the relion\_refine program. A helper program to find and optimize the 3D transformation matrices, relion\_localsym, was also implemented. To minimize artefacts in Fourier space, the edges of the masks should be kept soft, that is, they should gradually change from zero to one over multiple real-space pixels.

Our new implementation allows local symmetry averaging in both 3D classifications and 3D refinements, and works with 2D projection images as well as 3D images, such as sub-tomograms. At every step of

the expectation-maximization algorithm, local symmetry is applied according to the masks and transformations defined in the STAR file. This symmetrization is performed after the maximization step in real space. As such, the signal-to-noise gain that arises from the additional averaging is not considered when calculating the FSC between the two independent half-reconstructions in 3D auto-refinement, or when cal $culating \, the \, signal \, from \, the \, power \, of \, the \, reconstruction \, in \, 3D \, classification  tion. Therefore, in cases of local symmetry, it may be advantageous to increase the empirical regularization T-value, -tau2 fudge, to account for the expected gain in signal during the refinement. When doing so, care should be taken not to overfit the data by using a T-value that is too high. Calculation of a final FSC curve using Relion's post-processing program will lead to a realistic resolution estimate. However, care should be taken to use soft-edged masks for the local symmetrization, as the real-space mask operations may lead to artefacts in the Fourier-based resolution assessment.

Another note of caution considers the assumption that all the members of each of the local symmetry groups are identical. Since, by definition, the complex does not obey point group symmetry, this assumption can never be entirely true. At the very least, some of the subcomplex interfaces will be chemically different, whereas in worst-case scenarios biologically relevant conformational differences may exist within the members of each group. Imposing (local) symmetry on objects that are not identical will lead to a false impression of similarity in the output reconstruction. Besides minimizing artefacts in Fourier space, the soft edges of the local symmetry masks are also relevant here. Local symmetry is applied for all real-space pixels where the mask has values m > 0, but the symmetrized map will be calculated as (1-m) times the original reconstruction plus m times the average of the corresponding voxel in all members of the group. Thereby, mask values of m < 1 can be used to impose local symmetry only partially.

For the subcomplex reconstruction, masks around each of the symmetric regions were created and filtered to 25 Å with mask extension to 6 pixels and a soft-edge of 10 pixels (Extended Data Fig. 7e). The transformation operator for the three Euler angles and the three translations that relate one symmetric region with the other were calculated using relion\_localsym\_mpi in real space, followed by 3D refinement using a regularization T-value of 100. This improved the overall appearance of the map (Extended Data Fig. 7f).

### **Model building**

Model building was performed using Coot<sup>45,46</sup>. All models described were built as polyalanine chains except FANCL (see below).

A homology model for chicken FANCL (residues 1-373) was generated with I-TASSER<sup>47</sup> using *Drosophila* FANCL (PDB 3K1L)<sup>15</sup> as a template. The FANCL homology model was rigidly fitted into the EM map in Chimera using 'fit in map' tool<sup>48</sup> and then flexibly fitted using Jiggle fit in Coot. Densities corresponding to the ELF, URD and RING domains of FANCL were identified in the focused map of the base region, however, we were not able to orient the RING domain unambiguously in the density. The density for the ELF domain was well-defined in the focused map for the top region. There was only weak density for the URD domain and no density for the RING domain in FANCL  $_{\rm top}$  (Extended Data Figs. 2g, 5d). The ELF domain is next to the coiled-coil helix of FANCB, in agreement with FANCL crosslinks to the region C-terminal of the coiled-coil helix of FANCB (residues 439, 441, 454 and 460) and to the FAAP100  $\beta$ -propeller (residues 25, 180, 188, 262, 267 and 274). There are several crosslinks between FANCL URD-RING and the substrate-recognition  $module (FANCC, FANCE \, and \, FANCF) \, consistent \, with its \, placement \, within \, an extension \, and \, placement \, within \, and \, placement \, within \, an extension \, placement \, within \, and \, placement \, within \, an extension \, placement \, within \, and \, placement \, within \, an extension \, placement \, within \, and \, placement \, within \, an extension \, placement \, an extension \, an extensio$ the base of the complex.

The 'multi-body refinement map' for the middle region, the map from T=5 regularization, and the map for symmetric subcomplex were used to build the central region of FANCB-FAAP100 de novo by placing idealized helices and strands in Coot, and refining their fits with the real space refine zone tool. In agreement with their role as a scaffold, FANCB

and FAAP100 have 64 intermolecular crosslinks and they crosslink to four other subunits.

One of the two  $\beta$ -propellers in the top region was built de novo using the 'focused map' for the top region and this model was used to identify a structural homologue in DALI<sup>49</sup>. The  $\beta$ -propeller of Bardet–Biedl syndrome 1 protein (PDB 4VON)<sup>50</sup> was the top hit, and this was also a good fit for the other  $\beta$ -propeller. The sequence of this model was changed to polyalanine and used to build all four  $\beta$ -propellers after rigid fitting into the 'focused map' for the top or base region.

On the basis of the crosslinking patterns and secondary structure predictions, we assigned one of each of the pairs of  $\beta$ -propellers, and one helix of each of the coiled coils to FANCB (N-terminal region for  $\beta$ -propeller and residues ~390–429 for the helix) and the other to FAAP100 (N-terminal region for  $\beta$ -propeller and residues ~461–501 for the helix). These assignments agree with the hydrogen deuterium exchange experiments of B-100 versus B-L-100. The sequences of the predicted long helices of FANCB and FAAP100 were used in MARCOIL  $^{51,52}$  for assignment of heptad repeats (Extended Data Fig. 5e). Then, a FANCB–FAAP100 coiled-coil model was built in CCbuilder 2.0  $^{53}$  using the advanced mode (Extended Data Fig. 5f). This model was then fitted into the map and refined in Coot using real space refinement.

A homology model for chicken FANCG (residues 1–648) was generated in I-TASSER using TTC7B—hyccin complex (PDB 5DSE) and O-linked Glc-NAc transferase (PDB 1W3B) as the top two templates. The TPR region of the model (residues 153–432) was fitted into the focused map of the top region by rigid fitting in Chimera followed by Jiggle fit in Coot and refinement in Refmac. Additional helices were added towards the C-terminal region of FANCG using Coot.

A homology model for chicken FANCF (residues 117–343) was built in I-TASSER using the crystal structure of human FANCF C-terminal region (PDB 2IQC) $^{16}$  as one of the templates. Residues 117–191 and 215–343 were rigidly fitted into the focused map for the base region followed by Jiggle fit and real space refinement in Coot. In addition to the FANCF model, several short, idealized helices were also placed into the base. A crystal structure for human FANCE C-terminal region (PDB 2ILR) $^{54}$  was available but we could not fit it in the map, presumably due to conformational variation. The C terminus of FANCC is near FANCE (residues 155–190; Fig. 1d).

The above models were assembled together into models for top (FANCG,  $\beta$ -propellers and long helices of FANCB and FAAP100), middle and base (FANCL, FANCF,  $\beta$ -propellers and long helices of FANCB, FAAP100, unassigned helices) regions. These models were then further refined in Refmac<sup>55</sup> and Phenix iteratively<sup>56</sup>.

The models for FANCB<sub>top</sub>, FAAP100<sub>top</sub>, FANCB–FAAP100, FANCG and FANCL<sub>top</sub> built in the complete FA core complex map were rigidly fitted into the subcomplex map and refined in Refmac. All models and maps were visualized and rendered in UCSF Chimera<sup>48</sup> or ChimeraX<sup>57</sup>.

### Native mass spectrometry

Native mass spectrometry experiments were carried out on a Q-Exactive Plus UHMR modified to facilitate the transmission of high-energy species and adapted for membrane proteins<sup>58-60</sup>. The FA core complex at concentration of 9.5 µM was buffer exchanged using Bio-Spin 6 columns (BioRad). Typically, 2–3 μl of sample in 750 mM ammonium acetate were injected, using a 1.2 mm outer diameter, gold-coated borosilicate capillary (Harvard Apparatus). The following parameters were used for protein transmission: capillary voltage 1.2 kV, dessolvation voltage -300 V, source fragmentation 0 V, HCD energy 0 V, HCD pressure 4, EMR on, C-trap entrance lens tune offset 2, injection flatopole 8 V, inter flatopole lens 6 V, and bent flatopole 4 V. Threshold was set to 3. Data were analysed using Xcalibur 2.2 (Thermo Fisher), Masslynx 4.2 (Waters) and SUMMIT<sup>61</sup>. The formation of some of the subcomplexes observed may be the result of buffer exchange from 50 mM HEPES pH 8.0, ~500 mM NaCl and 1 mM TCEP into 750 mM ammonium acetate, which is often used in native mass spectrometry to improve resolution and generate subcomplexes to aid structure determination.

### HDX-MS

Deuterium exchange reactions were initiated by diluting the complexes in D $_2$ O (99.8% D $_2$ O ACROS, Sigma) to give a final D $_2$ O percentage of -95%. Deuterium labelling was generally carried out at 23 °C at four time points (3 s, 30 s, 300 s and 3,000 s) in triplicate. The labelling reaction was quenched by adding chilled 2.4% v/v formic acid in 2 M guanidinium hydrochloride and immediately frozen in liquid nitrogen. Samples were stored at -80 °C before analysis.

The quenched samples were rapidly thawed and subjected to proteolytic cleavage using pepsin followed by reverse-phase high performance liquid chromatography separation. The proteins flowed through an Enzymate BEH immobilized pepsin column, 2.1×30 mm, 5 µm (Waters) at 200 ul min<sup>-1</sup> for 2 min, and the resulting peptides were trapped and desalted on a 2.1 × 5 mm C18 trap column (Acquity BEH C18 Van-guard pre-column, 1.7 µm, Waters). Trapped peptides were eluted over 11 min using a 3-43% gradient of acetonitrile in 0.1% v/v formic acid at 40 μl min<sup>-1</sup> on to a reverse phase analytical column (Acquity UPLC BEH C18 column 1.7 µm, 100 mm × 1 mm (Waters)). The liquid chromatography elute was coupled to a SYNAPT G2-Si HDMS mass spectrometer (Waters) and data were acquired over a m/z of 300 to 2,000, using the standard electrospray ionization (ESI) source with lock mass calibration using [Glu1]-fibrino peptide B (50 fmol µl<sup>-1</sup>). The mass spectrometer was operated in ion mobility mode, at a source temperature of 80 °C and a spray voltage of 2.6 kV. Spectra were collected in positive-ion mode.

Peptide identification was performed with a non-deuterated sample using  $MS^E$  (Waters) to fragment peptides. An identical gradient of increasing acetonitrile in 0.1% v/v formic acid over 11 min was used and the resulting  $MS^E$  data were analysed using Protein Lynx Global Server software (Waters) with an MS tolerance of 5 ppm.

Mass analysis of the peptide centroids was performed using DynamX sotware (Waters). Only peptides with a score > 6.4 were considered. All peptides (deuterated and non-deuterated) were manually verified at every time point for the correct charge state, presence of overlapping peptides and correct retention time. Deuterium incorporation was not corrected for back-exchange and represents relative, rather than absolute changes in deuterium levels. Changes in H/D amide exchange in any peptide may be due to a single amide or several amides within that peptide.

# Crosslinking mass spectrometry of purified FA core complex

The purified FA core complex (50 mM HEPES pH 8.0, ~500 mM NaCl and 1 mM TCEP) at a concentration of 7.6  $\mu$ M was crosslinked with 100-fold molar ratio of disulfosuccinimidyl suberate (BS3) for 2 h on ice and the reaction was quenched with 50 mM NH $_4$ HCO $_3$  for 30 min at room temperature. The crosslinked samples were cold-acetone precipitated and resuspended in 8 M urea and 100 mM NH $_4$ HCO $_3$ . Peptides were reduced with 10 mM DTT and alkylated with 50 mM iodoacetamide. Following alkylation, proteins were digested with Lys-C (Pierce) at an enzyme-to-substrate ratio of 1:100 for 4 h at 22 °C and, after diluting the urea to 1.5 M with 100 mM NH $_4$ HCO $_3$  solution, further digestion with trypsin (Pierce) at an enzyme-to-substrate ratio of 1:20.

Digested peptides were eluted from StageTips and split into two, for parallel crosslink enrichment by strong cation-exchange chromatography (SCX) and size exclusion chromatography (SEC), and were dried in a vacuum concentrator (Eppendorf). For SCX, eluted peptides were dissolved in mobile phase A (30% acetonitrile (v/v), 10 mM KH $_2$ PO $_4$ , pH 3) before strong cation exchange chromatography (100 × 2.1 mm Poly Sulfoethyl A column; Poly LC). The separation of the digest used a nonlinear gradient citomobile phase B (30% acetonitrile (v/v), 10 mM KH $_2$ PO $_4$ , pH 3, 1M KCl) at a flow rate of 200  $\mu$ l min -1. Ten 1-min fractions in the high-salt range were collected and cleaned by StageTips, eluted and dried for subsequent liquid chromatography with tandem mass spectrometry (LC-MS/MS) analysis. For peptideSEC, peptides were fractionated on an ÄKTA Pure system (GE Healthcare) using a Superdex

Peptide 3.2/300 (GE Healthcare) at a flow rate of  $10\,\mu l\,min^{-1}$  using 30% (v/v) acetonitrile and 0.1% (v/v) trifluoroacetic acid as mobile phase. Five  $50-\mu l$  fractions were collected and dried for subsequent LC-MS/MS analysis.

Samples for analysis were resuspended in 0.1% v/v formic acid.1.6% v/v acetonitrile. LC-MS/MS analysis was conducted in duplicate for SEC fractions and triplicate for SCX fractions, performed on an Orbitrap Fusion Lumos Tribrid mass spectrometer (Thermo Fisher Scientific) coupled on-line with an Ultimate 3000 RSLCnano system (Dionex, Thermo Fisher Scientific). The sample was separated and ionized by a 50 cm EASY-Spray column (Thermo Fisher Scientific). Mobile phase A consisted of 0.1% (v/v) formic acid and mobile phase B of 80% v/v acetonitrile with 0.1% v/v formic acid. Flow-rate of 0.3 µl min<sup>-1</sup> using gradients optimized for each chromatographic fraction from offline fractionation ranging from 2% mobile phase B to 45% mobile phase B over 90 min, followed by a linear increase to 55% and 95% mobile phase B in 2.5 min, respectively. The MS data were acquired in data-dependent mode using the top-speed setting with a three second cycle time. For every cycle, the full scan mass spectrum was recorded in the Orbitrap at a resolution of 120,000 in the range of 400 to 1,600 m/z. lons with a precursor charge state between 3+ and 7+ were isolated and fragmented. Fragmentation by higher-energy collisional dissociation (HCD) employed a decision tree logic with optimized collision energies<sup>63</sup>. The fragmentation spectra were then recorded in the Orbitrap with a resolution of 50,000. Dynamic exclusion was enabled with single repeat count and 60-s exclusion duration.

A recalibration of the precursor m/z was conducted based on high-confidence (<1% false discovery rate (FDR)) linear peptide identifications  $^{64}$ . The recalibrated peak lists were searched against the sequences and the reversed sequences (as decoys) of crosslinked peptides using the Xi software suite (v.1.6.746) $^{65}$  (https://github.com/Rappsilber-Laboratory/XiSearch) for identification. The following parameters were applied for the search: MS1 accuracy = 3 ppm; MS2 accuracy = 10 ppm; enzyme = trypsin (with full tryptic specificity) allowing up to four missed cleavages; crosslinker = BS3 with an assumed reaction specificity for lysine, serine, threonine, tyrosine and protein N termini; fixed modifications = carbamidomethylation on cysteine; variable modifications = oxidation on methionine, hydrolyzed/aminolyzed BS3 from reaction with ammonia or water on a free crosslinker end. The identified candidates were filtered to 1% FDR on link level using XiFDR v.1.1.26.58 $^{66}$ .

### Pull down assay

The purified subcomplexes A-G-B-L-100 and C-E-F (Strep II tag cleaved by 3C protease) were mixed in a 1:1 molar ratio at concentrations of 1.4  $\mu$ M each, for 1h at 4 °C. A 20- $\mu$ l reaction was incubated with 15  $\mu$ l of StrepTactin beads (GE Healthcare) equilibrated in 50 mM HEPES pH 8.0, 300 mM NaCl and 1 mM TCEP. The flow through was collected (unbound fraction) and the beads were washed three times with equilibration buffer. The unbound and bound fractions were analysed on SDS-PAGE.

### **Reporting summary**

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

### **Data availability**

Cryo-EM maps generated during this study have been deposited in the Electron Microscopy Data Bank with accession codes EMD-10290 (FA core complex consensus), EMD-10291 (focused classification top region), EMD-10292 (focused classification middle region), EMD-10293 (focused classification base region) and EMD-10294 (subcomplex). Models generated during this study have been deposited in the Protein Data Bank (PDB) with accession codes 6SRI (FA core complex) and 6SRS (subcomplex). Native mass spectrometry data are available from figshare at https://doi.org/10.6084/m9.figshare.9692192. Crosslinking mass spectrometry data have been deposited in the PRIDE database

with accession code PXD014282. All other data are available from the authors upon reasonable request.

- Sari, D. et al. The MultiBac baculovirus/insect cell expression vector system for producing complex protein biologics. Adv. Exp. Med. Biol. 896, 199–215 (2016).
- Weissmann, F. et al. biGBac enables rapid gene assembly for the expression of large multisubunit protein complexes. Proc. Natl Acad. Sci. USA 113, E2564–E2569 (2016).
- Hill, C. H. et al. Activation of the endonuclease that defines mRNA 3' ends requires incorporation into an 8-subunit core cleavage and polyadenylation factor complex. Mol. Cell 73, 1217–1231.e11 (2019).
- Sato, K., Toda, K., Ishiai, M., Takata, M. & Kurumizaka, H. DNA robustly stimulates FANCD2 monoubiquitylation in the complex with FANCI. *Nucleic Acids Res.* 40, 4553–4561 (2012).
- Russo, C. J. & Passmore, L. A. Ultrastable gold substrates for electron cryomicroscopy. Science 346, 1377–1380 (2014).
- Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. J. Struct. Biol. 180, 519–530 (2012).
- Fernandez-Leiro, R. & Scheres, S. H. W. A pipeline approach to single-particle processing in RELION. Acta Crystallogr. D 73, 496–502 (2017).
- Zivanov, J. et al. New tools for automated high-resolution cryo-EM structure determination in RELION-3. eLife 7, e42166 (2018).
- Tang, G. et al. EMAN2: an extensible image processing suite for electron microscopy. J. Struct. Biol. 157, 38–46 (2007).
- Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. Nat. Methods 14, 331–332 (2017).
- Zhang, K. Gctf: Real-time CTF determination and correction. J. Struct. Biol. 193, 1–12 (2016).
- Scheres, S. H. A Bayesian view on cryo-EM structure determination. J. Mol. Biol. 415, 406–418 (2012).
- Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. Nat. Methods 11, 63–65 (2014).
- García-Nafría, J., Lee, Y., Bai, X., Carpenter, B. & Tate, C. G. Cryo-EM structure of the adenosine A<sub>2A</sub> receptor coupled to an engineered heterotrimeric G protein. eLife 7,
- Stewart, P. L., Burnett, R. M., Cyrklaff, M. & Fuller, S. D. Image reconstruction reveals the complex molecular organization of adenovirus. Cell 67, 145–154 (1991).
- He, J., Schmid, M. F., Zhou, Z. H., Rixon, F. & Chiu, W. Finding and using local symmetry in identifying lower domain movements in hexon subunits of the herpes simplex virus type 1 B capsid. J. Mol. Biol. 309, 903–914 (2001).
- Rossmann, M. G. & Blow, D. M. Detection of sub-units within crystallographic asymmetric unit. Acta Crystallogr. D 15, 24–31 (1962).
- Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. Acta Crystallogr. D 66, 486–501 (2010).
- Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. Acta Crystallogr. D 60, 2126–2132 (2004).
- Yang, J. et al. The I-TASSER suite: protein structure and function prediction. Nat. Methods 12, 7–8 (2015).
- 48. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
- Holm, L. & Sander, C. Dali: a network tool for protein structure comparison. Trends Biochem. Sci. 20, 478–480 (1995).
- Mourão, A., Nager, A. R., Nachury, M. V. & Lorentzen, E. Structural basis for membrane targeting of the BBSome by ARL6. Nat. Struct. Mol. Biol. 21, 1035–1041 (2014).
- Delorenzi, M. & Speed, T. An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. *Bioinformatics* 18, 617–625 (2002).
- Zimmermann, L. et al. A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. J. Mol. Biol. 430, 2237–2243 (2018).
- Wood, C. W. & Woolfson, D. N. CCBuilder 2.0: powerful and accessible coiled-coil modeling. *Protein Sci.* 27, 103–111 (2018).
- Nookala, R. K., Hussain, S. & Pellegrini, L. Insights into Fanconi anaemia from the structure of human FANCE. Nucleic Acids Res. 35, 1638–1648 (2007).
- Murshudov, G. N. et al. REFMAC5 for the refinement of macromolecular crystal structures. Acta Crystallogr. D 67, 355–367 (2011).
- Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr. D 66: 213–221 (2010).
- Goddard, T. D. et al. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. Protein Sci. 27, 14–25 (2018).
- van de Waterbeemd, M. et al. High-fidelity mass analysis unveils heterogeneity in intact ribosomal particles. Nat. Methods 14, 283–286 (2017).
- Rose, R. J., Damoc, E., Denisov, E., Makarov, A. & Heck, A. J. High-sensitivity Orbitrap mass analysis of intact macromolecular assemblies. *Nat. Methods* 9, 1084–1086 (2012).
- Gault, J. et al. High-resolution mass spectrometry of small molecules bound to membrane proteins. Nat. Methods 13, 333–336 (2016).
- Taverner, T. et al. Subunit architecture of intact protein complexes from mass spectrometry and homology modeling. Acc. Chem. Res. 41, 617–627 (2008).
- Chen, Z. A. et al. Architecture of the RNA polymerase II-TFIIF complex revealed by crosslinking and mass spectrometry. EMBO J. 29, 717–726 (2010).
- Kolbowski, L., Mendes, M. L. & Rappsilber, J. Optimizing the parameters governing the fragmentation of cross-linked peptides in a tribrid mass spectrometer. *Anal. Chem.* 89, 5311–5318 (2017).
- Lenz, S., Giese, S. H., Fischer, L. & Rappsilber, J. In-search assignment of monoisotopic peaks improves the identification of cross-linked peptides. *J. Proteome Res.* 17, 3923–3931 (2018).
- Giese, S. H., Fischer, L. & Rappsilber, J. A study into the collision-induced dissociation (CID) behavior of cross-linked peptides. Mol. Cell. Proteomics 15, 1094–1104 (2016).
- Fischer, L. & Rappsilber, J. Quirks of error estimation in cross-linking/mass spectrometry. Anal. Chem. 89, 3829–3833 (2017).

- Naydenova, K. & Russo, C. J. Measuring the effects of particle orientation to improve the efficiency of electron cryomicroscopy. Nat. Commun. 8, 629 (2017).
- Buetow, L. & Huang, D. T. Structural insights into the catalysis and regulation of E3 ubiquitin ligases. Nat. Rev. Mol. Cell Biol. 17, 626–642 (2016).
- Knipscheer, P. & Sixma, T. K. Protein-protein interactions regulate Ubl conjugation. Curr. Opin. Struct. Biol. 17, 665–673 (2007).
- Metzger, M. B., Pruneda, J. N., Klevit, R. E. & Weissman, A. M. RING-type E3 ligases: master manipulators of E2 ubiquitin-conjugating enzymes and ubiquitination. *Biochim. Biophys.* Acta 1843, 47–60 (2014).
- Linares, L. K., Hengstermann, A., Ciechanover, A., Müller, S. & Scheffner, M. HdmX stimulates Hdm2-mediated ubiquitination and degradation of p53. *Proc. Natl Acad. Sci.* USA 100, 12009–12014 (2003).
- Alpi, A. F., Pace, P. E., Babu, M. M. & Patel, K. J. Mechanistic insight into site-restricted monoubiquitination of FANCD2 by Ube2t, FANCL, and FANCI. Mol. Cell 32, 767–777 (2008).

Acknowledgements We thank T. Nakane, J. Zivanov, C. Lau, A. Carter, P. Emsley, G. Murshudov, D. Malinverni and M. Babu for advice and discussions; K. Naydenova, B. Santhanam, G. Dornan, D. Briant, A. Casañal, A. Kumar, M. Carminati, A. Kelley and members of the Passmore laboratory for assistance; G. Cannone, C. Savva and the LMB EM facility, J. Grimmett and T. Darling (LMB scientific computation) for support and J. Shi for assistance with baculovirus. This work was supported by the Medical Research Council, as part of United Kingdom Research and Innovation, MRC file reference number MC\_U105192715 (L.A.P.); Deutsche

Forschungsgemeinschaft (DFG, 329673113) (J.R.); the Wellcome Trust through a Senior Research Fellowship to J.R. (103139); and the European Research Council grant number 695511-ENABLE (C.V.R). The Wellcome Centre for Cell Biology is supported by core funding from the Wellcome Trust (203149). We acknowledge Diamond Light Source for access to eBIC (proposals EM18091 and EM17434) funded by the Wellcome Trust, MRC and Biotechnology and Biological Sciences Research Council.

**Author contributions** S.S., E.R. and P.A. designed protein expression and purification schemes, performed ubiquitination assays and performed cryo-EM, 3D reconstruction and modelling; S.S. and D.S.C. performed native mass spectrometry; F.O'R. and G.D. performed crosslinking mass spectrometry; S.M. performed hydrogen-deuterium exchange-mass spectrometry; S.H. and S.H.W.S. developed the local symmetry algorithm; C.H.H., C.J.R. and L.A.P. collected cryo-EM data; L.A.P., C.V.R., J.M.S., J.R., S.H.W.S. and K.J.P. supervised the research; L.A.P. conceived the project; and S.S., P.A. and L.A.P. wrote the paper with contributions from all authors.

Competing interests The authors declare no competing interests.

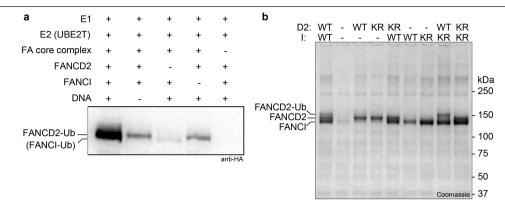
### Additional information

**Supplementary information** is available for this paper at https://doi.org/10.1038/s41586-019-1703-4

Correspondence and requests for materials should be addressed to L.A.P.

Peer review information *Nature* thanks Andrew Deans, Xiaodong Zhang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

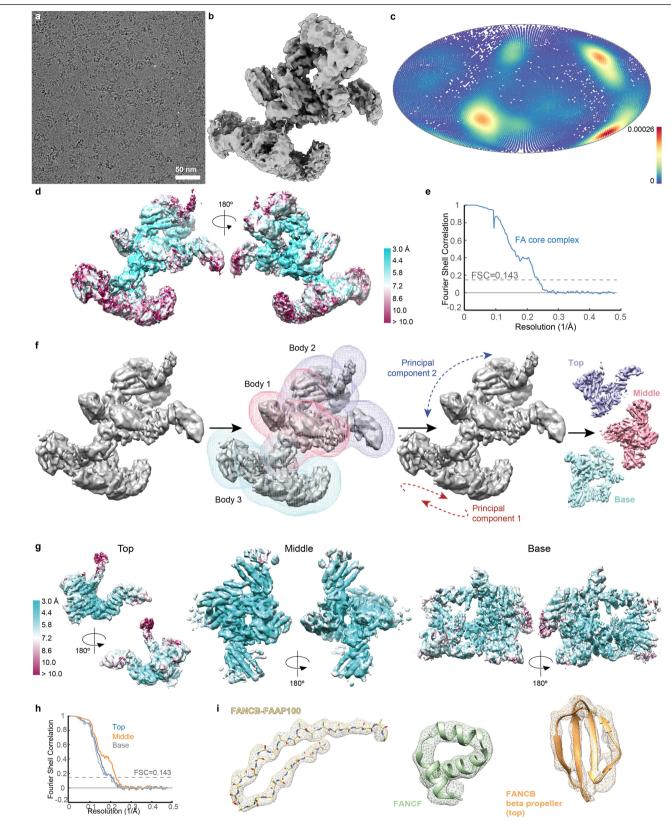
Reprints and permissions information is available at http://www.nature.com/reprints.



### Extended Data Fig. 1 | Recombinant FA core complex activity.

**a**, Ubiquitination assay analysed by western blot with HA antibody to detect HA-tagged ubiquitin. The migration positions of monoubiquitinated FANCD2 and FANCI are indicated but FANCI is not substantially modified. **b**, Ubiquitination assay analysed by Coomassie-stained SDS-PAGE to show specific monoubiquitination of FANCD2 K563 by recombinant FA core complex. Wild type (WT), FANCD2(K563R) and FANCI(K525R) (KR) were analysed. A native FA core complex purified from chicken DT40 cells monoubiquinates FANCD2 but does not efficiently monoubiquitinate FANCI<sup>6</sup>.

Therefore, the purified recombinant complex faithfully recapitulates the properties of the native chicken complex. Notably, a purified human complex also did not efficiently monoubiquitinate FANCI, although it did efficiently monoubiquitinate FANCD2<sup>11</sup>. The asymmetry in the FA core complex (see below) reflects this asymmetry in its activity on FANCD2–FANCI. An additional factor or post-translational modification may be required for activation of FANCI monoubiquitination. The ubiquitination assays were repeated at least two times independently with similar results. For gel source data, see Supplementary Fig. 1.

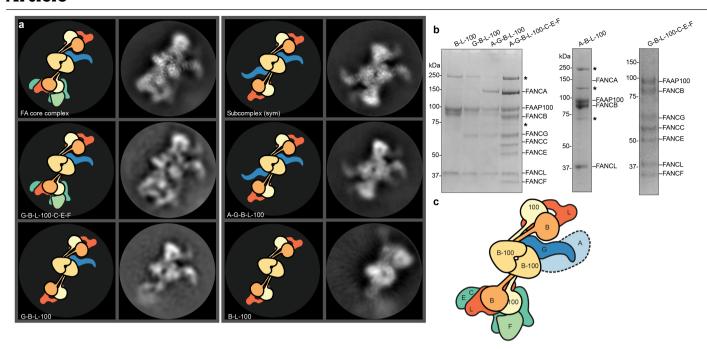


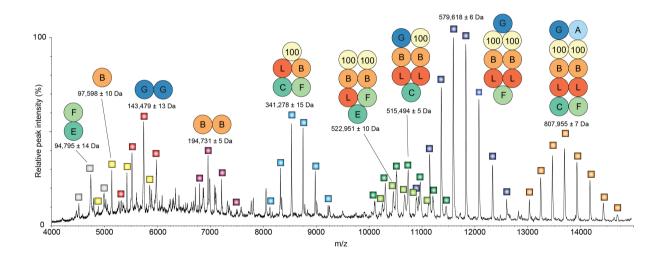
**Extended Data Fig. 2** | See next page for caption.

Extended Data Fig. 2 | Cryo-EM reconstruction of FA core complex, multibody refinement and assessment of 3D reconstructions after focused refinement. a, Representative raw micrograph of FA core complex. b, Overall 3D reconstruction of the FA core complex. c, Angular distribution density plot of particles used in the 3D reconstruction of the FA core complex. Every point is a particle orientation and the colour scale represents the normalized density of views around this point. The colour scale runs from 0 (low, blue) to 0.00026 (high, red). The efficiency of orientation distribution 67, E<sub>OD</sub>, was 0.79. d, Estimated local resolution map for FA core complex. e, FSC plot for gold

standard refinement. f, Multibody refinement of the FA core complex using

three masks (body 1, body 2 and body 3) shown in pink, purple and cyan, respectively. The motions are shown in Supplementary Videos 1 and 2.  ${\bf g}$ , Local resolution maps for reconstructions of top, middle and base regions of FA core complex. The middle region did not substantially change between multi-body refinement and particle subtraction followed by focused classification. The resolution of the base and top regions improved after particle subtraction and focused classification and refinement.  ${\bf h}$ , FSC plot for gold standard refinements.  ${\bf i}$ , Representative density for  ${\boldsymbol \beta}$ -strand and  ${\boldsymbol \alpha}$ -helical regions. FANCB–FAAP100 is in the middle region which is better defined than more peripheral regions, including FANCF.





d

Protein	Expected	Measured	Phosphorylation
	mass (Da)	mass (Da)*	sites#
FANCA	159,828	159,888 ± 3	n.d.
FANCB	97,047	97,598 ± 10	n.d.
FANCC-SII <sup>†</sup>	67,476	67,383 ± 2	n.d.
FANCE	55,962	56,089 ± 1	S135, S250
FANCF	38,322	38,364 ± 2	n.d.
FANCG	71,786	71,825 ± 54	n.d.
FANCL	42,638	42,609 ± 11	n.d.
FAAP100	95,406	95,883 ± 91	S617, S619, S621,
			T664, S670

<sup>\*</sup> Measured mass in native MS ± standard deviation

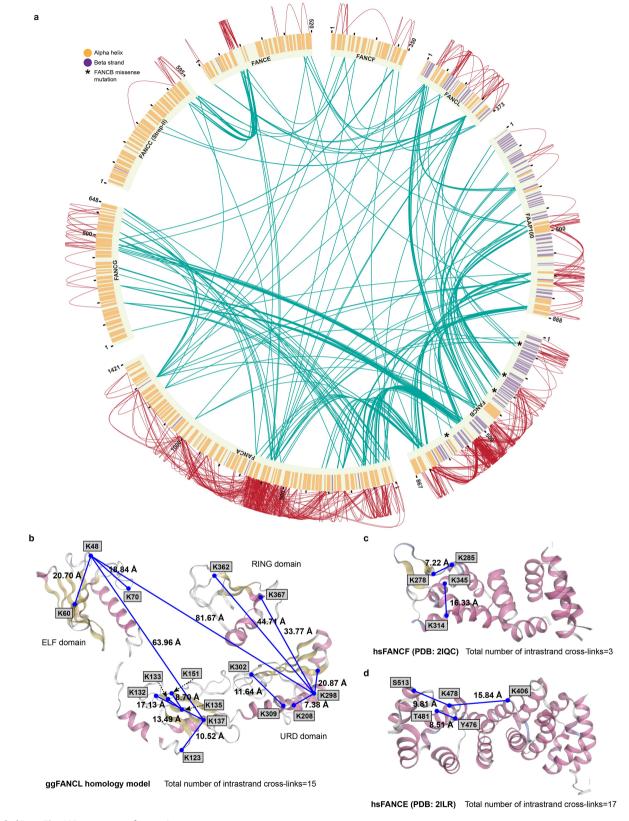
**Extended Data Fig. 3** | See next page for caption.

<sup>#</sup> Phosphorylation sites on purified protein were identified by tandem mass spectrometry
† Mass of FANCC includes the SII tag

n.d., none detected

Extended Data Fig. 3 | Subunit assignment and arrangement in FA core  $\textbf{complex.}\,\textbf{a-c}, \textbf{Complexes lacking specific subunits were purified and analysed}$ by cryo-EM. a, Major 2D class averages identified for subcomplexes, compared with those from FA core complex (A-G-B-L-100-C-E-F). Cartoons are shown to depict the subunits visible in the class averages. The symmetric subcomplex identified in the FA core complex preparation is indicated (sym). The A-G-B-L-100 complex (lacking the substrate-recognition module) has similar symmetric 2D classes. Native mass spectrometry revealed a non-uniform subunit stoichiometry. Thus it is likely that the asymmetric assembly represents the complete FA core complex, while the symmetric structure is a subcomplex that co-purifies with the intact complex. The 2D class average of a complex lacking FANCA (G-B-L-100-C-E-F) appeared similar to the complete FA core complex with no obvious missing density. FANCG is probably the partially disordered arm that extends from the central part of the complex since this was missing when FANCG was not present in the complex. b, Coomassie-stained  $SDS-PAGE\ analysis\ of\ purified\ subcomplexes.\ Asterisks\ indicate\ contaminant$  $proteins.\,FANCA\,did\,not\,co\text{-}purify\,with\,the\,A\text{-}B\text{-}L\text{-}100\,complex\,but\,its\,migration}$ position is indicated on the gel. The purifications were repeated at least two

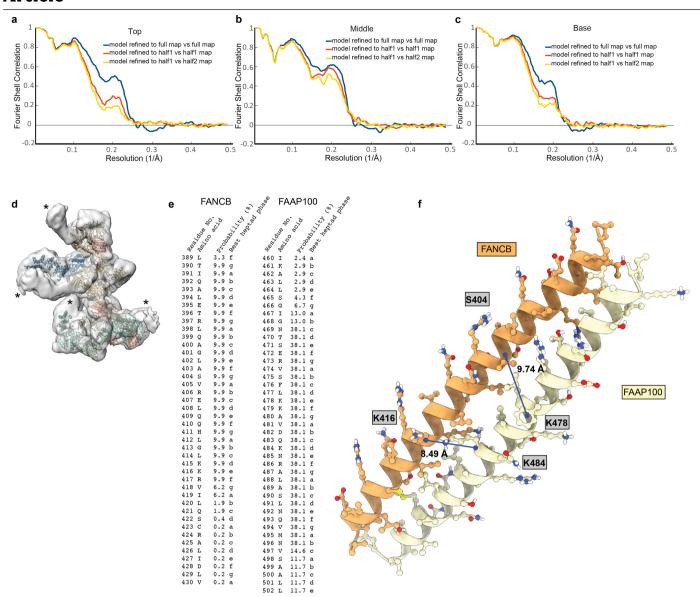
times independently with similar results. For gel source data, see Supplementary Fig. 1.c, Cartoon of FA core complex with subunits labelled. d, Native mass spectrum of recombinant FA core complex showing masses and subunit composition of assigned peak series. We dissociated the FA core complex into subcomplexes during ionization and these species were detected  $by\,mass\,spectrometry.\,Computational\,analyses\,then\,revealed\,the\,proteins$ present in each of the peaks. The standard deviation in fitting the identified peaks to the charge series is given as the ± error in the measured mass for a given single measurement. This is the error in the fit and not the error in the mass measurement, which is probably an order of magnitude higher due to, for instance, solvation or adduct effects, or heterogeneous post-translational modifications. Hence, the error gives a rough measure of the accuracy of peak assignment, which is impacted by the broadness, symmetry and signal-to-noise ratio of each peak. e, Molecular masses of FA core complex subunits. The expected and measured masses are given, along with phosphorylation sites  $detected \, by \, mass \, spectrometry. \, Native \, mass \, spectrometry \, was \, repeated \, three \,$ times with similar results.



 $\textbf{Extended Data Fig. 4} \ | \ See \ next \ page \ for \ caption.$ 

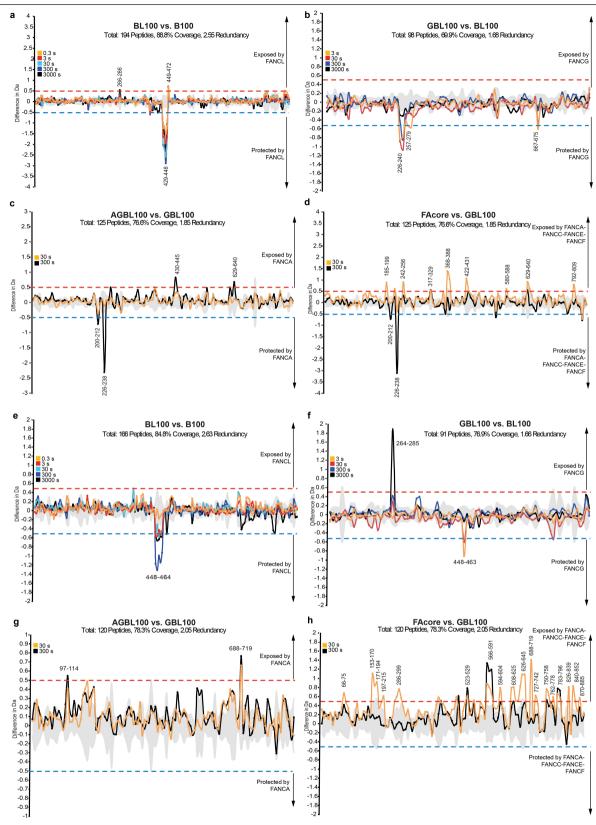
Extended Data Fig. 4 | Crosslinking mass spectrometry. a, Crosslinking mass spectrometry revealed 834 crosslinks (1% FDR) between residues that are in close proximity. Intermolecular crosslinks are coloured in green; intramolecular crosslinks, red; predicted  $\alpha$ -helices, orange; predicted  $\beta$ -strands, purple. Asterisks on FANCB mark missense mutations from the Fanconi Anaemia Mutation Database (http://www2.rockefeller.edu/fanconi/) including L43S in a predicted  $\beta$ -strand, P23OS at the N terminus of a predicted helix, and L329P in the middle of a predicted  $\beta$ -strand, all in the  $\beta$ -propeller; and L676P, which is predicted to disrupt an  $\alpha$ -helix in the C-terminal dimerization domain. b-d, Validation of crosslinking of FA core complex. Crosslinks were mapped onto a homology model of chicken FANCL (b), human FANCF (c) and human FANCE (d). All intramolecular crosslinks within a domain are consistent with the

maximum crosslinker length (30 Å between the two C $\alpha$ ). Crosslinks between the domains in FANCL are not consistent with the domain arrangement in the crystal structure of FANCL because there is flexibility or changes in the orientation between the domains in the FA core complex. There are two different conformations of FANCL in the structure: FANCL base is fully ordered, whereas only the ELF domain is visible in FANCL top. Mapping the crosslinks onto FANCL from the base (constrained, with all three domains visible) reveals that the distances for some crosslinks between domains are too large to be consistent with the FANCL conformation in the base. By contrast, for FANCL in the top, where only the ELF domain is ordered, the URD and RING domains are likely to be conformationally flexible. Since the URD and RING domains cannot be modelled for FANCL top, it is not possible to validate these interdomain crosslinks.



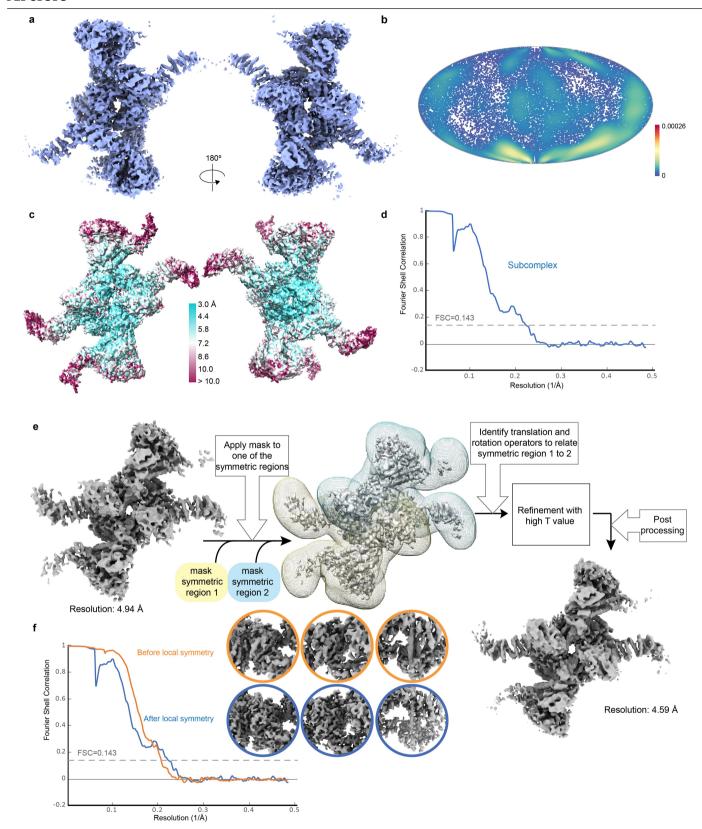
Extended Data Fig. 5 | Assessment of model fit in maps and modelling of coiled coils. a-c, FSC plots of maps versus model for top (a), middle (b) and base (c) regions. d, Low-resolution cryo-EM map of FA core complex (transparent surface) with models placed in the map. Asterisks represent density that was not visible at high resolution and was not interpreted. This may

represent FANCA and additional parts of the substrate-recognition module. e, MARCOIL  $^{\rm 51}$  prediction for the best heptad phase in the long helices of FANCB and FAAP100. f, Predicted coiled-coil model by CCbuilder 2.0  $^{\rm 53}$  for the FANCB and FAAP100 long helices. Crosslinks detected for these helices in the FA core complex crosslinking mass spectrometry are indicated with blue lines.



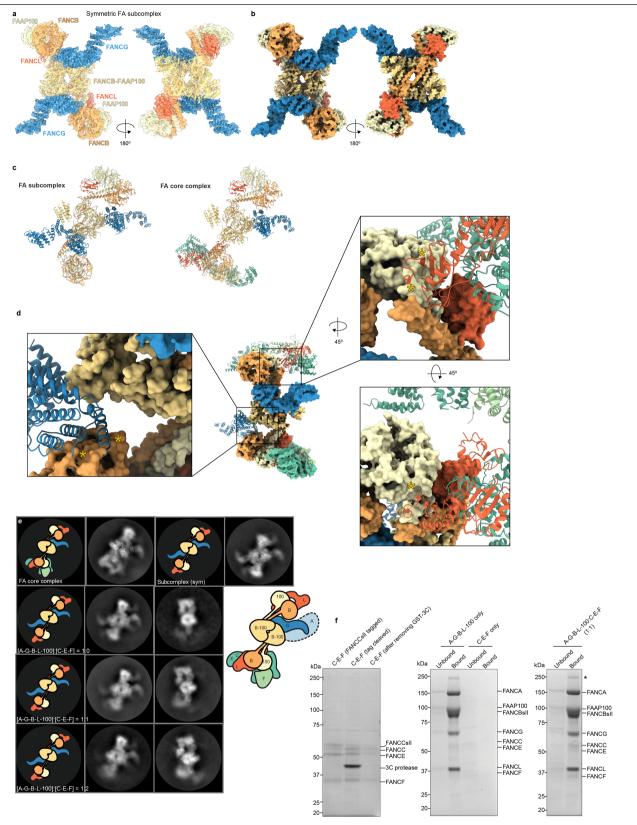
**Extended Data Fig. 6 | HDX-MS on FANCB and FAAP100.** a-d, Difference plots for FANCB showing peptides that are protected (negative) or exposed (positive) upon binding of additional subunit(s) for B-L-100 vs B-100 (a), G-B-L-100 vs B-L-100 (b), A-G-B-L-100 vs G-B-L-100 (c) and FA core complex vs. G-B-L-100 (d). Exchange of hydrogens in FANCB residues 429–448 was protected after interaction with FANCL, consistent with FANCL being located next to the coiled coil. e-h, Difference plots for FAAP100 showing peptides that are protected (negative) or exposed (positive) upon binding of additional

subunit(s) for B-L-100 vs B-100 ( $\mathbf{e}$ ), G-B-L-100 vs B-L-100 ( $\mathbf{f}$ ), A-G-B-L-100 vs. G-B-L-100 ( $\mathbf{g}$ ) and FA core complex vs G-B-L-100 ( $\mathbf{h}$ ). Exchange of hydrogens in FAAP100 residues 448–464 was protected after interaction with FANCL, consistent with FANCL being located next to the coiled coil. For difference plots, triplicate data from four independent colour-coded time points are shown. The significance threshold is indicated by dashed lines. Grey shading indicates the standard deviation of all charge states and replicates per peptide. Sequence coverage is shown in the Supplementary Information.



**Extended Data Fig. 7** | **3D reconstruction of symmetric FA subcomplex and local symmetry refinement. a**, Overall 3D reconstruction of the symmetric FA subcomplex. **b**, Angular distribution plot of particles used in the 3D reconstruction of the symmetric FA subcomplex. Every point is a particle orientation and the colour scale represents the normalized density of views around this point. The colour scale runs from 0 (low, blue) to 0.00026 (high, red). The efficiency of orientation distribution  $^{67}$ ,  $E_{\rm OD}$ , was 0.65. **c**, Estimated local

resolution map for symmetric FA subcomplex.  ${\bf d}$ , FSC plot for gold standard refinement.  ${\bf e}$ , Local symmetry pipeline for reconstruction of the symmetric FA subcomplex (see Methods). This reconstruction could not be improved with C2 symmetry, probably because of local flexibility.  ${\bf f}$ , FSC plot for gold standard refinement shown for the subcomplex reconstruction before and after local symmetry refinement. The circular panels show representative densities before and after local symmetry refinement.



**Extended Data Fig. 8** | See next page for caption.

Extended Data Fig. 8 | Model for FA subcomplex. a, Model of FA subcomplex shown as cartoon representations of subunits fit into the cryo-EM map. b, Model of FA subcomplex shown as a surface representation of the combined models. Two views are shown down the two-fold symmetric axis. c, Comparison of FA subcomplex and complete FA core complex in the same orientations. Both models are shown in cartoon representation. Subunits are coloured as in Fig. 1e. d, Modelling of a fully symmetric FA core complex containing two copies of every subunit. Left, the second copy of FANCG (cartoon) from the FA subcomplex was modelled onto the structure of the FA core complex (surface representation). This second FANCG clashes with the FANCB  $\beta$ -propeller in the base (asterisks). Thus, it is likely that upon binding of the substrate-recognition module, rearrangement of the β-propellers of FANCB and FAAP100 prevents binding of a second copy of FANCG. (Right) A second copy of substraterecognition module (FANCC-FANCE-FANCF; cartoon) is modelled in the top region of the FA core complex by combining the models of FANCC-FANCE- $FANCF \, and \, FANCL_{base} from \, FA \, core \, complex \, followed \, by \, superimposing \,$  $FANCL_{base}\ on\ FANCL_{top}.\ There \ is\ a\ clash\ (asterisks)\ between\ the\ modelled\ FANCL$ (cartoon) and FAAP100 β-propeller (surface representation). These data suggest that a fully symmetric complex does not readily form. In agreement with this, there was no evidence for any classes containing two copies of C-E-F in any of our EM analyses of the FA core complex. e, f, The symmetric FA subcomplex

(A-G-B-L-100) does not readily associate with purified substrate-recognition module (C-E-F) to form the asymmetric FA core complex. e, The 2D class averages of A-G-B-L-100 mixed with C-E-F compared with complete FA core complex, FA subcomplex and A-G-B-L-100 subcomplex. A-G-B-L-100 was mixed with C-E-F in molar ratios of 1:1 and 1:2 for 1 h at 4 °C before cryo-plunging. Only the symmetric A-G-B-L-100 subcomplex was observed and there was no additional density for C-E-F. Panels for FA core complex, subcomplex and A-G-B-L-100 are replicated from Extended Data Fig. 3a. f, Pull-down assay of C-E-F using tagged A-G-B-L-100 (Strep II tagged). Left, the Coomassie Blue-stained gel of purified C-E-F (with Strep II tag, after 3C cleavage of tag and after removal of 3C protease). Tagged A-G-B-L-100 was immobilized on StrepTactin resin and incubated with purified C-E-F at a 1:1 molar ratio. After washing, only a small amount of C-E-F remains bound to the beads. Negative controls (A-G-B-L-100 only and C-E-F only) are shown in the middle panel. Asterisk indicates a contaminant protein. The pulldown experiment was repeated two times independently with similar results. Since C-E-F does not efficiently bind A-G- $B-L-100, these \, experiments \, suggest \, that \, these \, species \, are \, unlikely \, to \, be \, in \, an experiment \, suggest \, that \, these \, species \, are \, unlikely \, to \, be \, in \, an experiment \, suggest \, that \, these \, species \, are \, unlikely \, to \, be \, in \, an experiment \, suggest \, that \, these \, species \, are \, unlikely \, to \, be \, in \, an experiment \, suggest \, that \, these \, species \, are \, unlikely \, to \, be \, in \, an experiment \, suggest \, that \, the \, species \, are \, unlikely \, to \, be \, in \, an experiment \, suggest \, that \, the \, species \, are \, unlikely \, to \, be \, in \, an experiment \, suggest \, that \, the \, species \, are \, unlikely \, to \, be \, in \, an experiment \, suggest \, that \, the \, species \, are \, unlikely \, to \, be \, in \, an experiment \, species \, are \, unlikely \, to \, be \, in \, an experiment \, species \, are \, unlikely \, to \, be \, in \, an experiment \, species \, are \, unlikely \, to \, be \, in \, an experiment \, species \, are \, unlikely \, to \, be \, in \, an experiment \, an$ equilibrium in solution. Previous genetic and biochemical data show that FANCE, FANCE and FANCF are important for monoubiquitination of the FANCD2-FANCI substrate. Together, these data provide evidence that the asymmetric complex is the relevant, functional, physiological entity.

Extended Data Fig. 9 | Structural comparison of E3 ligases. a, There is a strong precedence for dimerization of RING/U-box domain E3 ubiquitin ligases<sup>68-70</sup>. RING/U-box E3s exist both as homo- and heterodimeric complexes. for example, Rad18-Rad18, CHIP-CHIP, RNF8-RNF8, BRCA1-BARD1, RING1b-BMI1 and Hdm2-Hdmx<sup>17-21,71</sup>. Structures of homo- and heterodimeric RING/Ubox E3 ligases are shown here with the RING/U-box in orange. Surprisingly, these E3s display functional and structural asymmetry: in all the dimers listed above, only one protomer binds to an E2 enzyme. The homodimeric CHIP E3 ligase has a strikingly asymmetric structure that clearly demonstrates why only one U-box binds E2 enzyme19. The FANCL RING subunit is also an asymmetric dimer within the FA core complex and it is possible that only one of these binds E2. However, unlike the smaller E3 s, the FANCL RING fingers are not near each other. Together, this suggests that asymmetric dimerization may be a general feature of RING E3s. b, Comparison of FA core complex with cullin-RING ubiquitin ligases (CRLs). Many large complexes are predominantly helical suggesting that α-helices are commonly used as building blocks for complexes. In addition,  $\beta$ -propellers often mediate protein-protein interactions. The CRL complexes and FA core complex are long and extended with substraterecognition (green), scaffold (yellow) and RING (orange) subunits residing in three different regions of the structure. However, the structural details of these complexes differ. Interestingly, the activities of some multisubunit RINGcontaining E3 ligases including APC/C and CRL complexes are stimulated by dimerization<sup>22,23</sup>. Thus, dimerization may underpin physiological ubiquitination activity in many E3 ligases. c, d, Ubiquitin discharge assay, in

which free lysine is used instead of the FANCD2-FANCI substrate. In these experiments, the FA core complex is incubated with E1, E2, ubiquitin and free lysine. If FANCL is active without a substrate, ubiquitin will be conjugated to lysine, resulting in a shift in its molecular weight; however, if substrate binding is required to activate the E3 ligase activity (for example, through allosteric changes), this will not occur. Coomassie gels of reaction products were run in non-reducing (c) and reducing (100 mM DTT) conditions (d). Ubiquitin is transferred to free lysine as shown by the increase in molecular weight of ubiquitin as well as a decrease in intensity of the E2-ubiquitin band when compared to the lane containing no free lysine. Thus, substrate binding is not required for activity. Reducing conditions do not eliminate the UBE2Tubiquitin conjugate, as previously shown<sup>72</sup>. Additionally, DNA is not required for FA core complex E3 ligase activity on free lysines, suggesting that DNA activates the substrate, not the E3. The ubiquitin discharge assays were repeated three times independently with similar results  $(\mathbf{c}, \mathbf{d})$ . e, Distributions of patient mutations are indicated on the FA core complex by

heat map colouring of subunits and in percentage. **f**, Ubiquitination assay using several subcomplexes (Extended Data Fig. 3b) and the full FA core complex, analysed by western blot with HA antibody to detect HA-tagged ubiquitin. The migration positions of monoubiquitinated FANCD 2 and FANCI are indicated but FANCI is not substantially modified, as in Extended Data Fig. 1a. All complexes have similar activities but isolated FANCL is less active. This assay was repeated at least two times independently with similar results. For gel source data, see Supplementary Fig. 1.

# Extended Data Table 1 | Cryo-EM data collection, refinement and validation statistics

	FA core complex (EMD-10290: consensus, EMD-10291: top, EMD-10292: middle, EMD-10293: base) (PDB 6SRI)	Subcomplex (EMD-10294) (PDB 6SRS)
Data collection and processing		
Magnification	75,000 X	75,000 X
Voltage (keV)	300	300
Electron exposure (e–/Å <sup>2</sup> )	40	40
Defocus range (μm)	-1.8 to -4.0	-1.8 to -4.0
Pixel size (Å)	1.040 (LMB)	1.040 (LMB)
	1.085 (eBIC)	1.085 (eBIC)
Symmetry imposed	C1	C1
Initial particle images (no.)	1,947,765	1,947,765
Final particle images (no.)	169,000	49,423
Map resolution (Å)	4.2 (consensus), 4.5 (top),	4.6
•	4.4 (middle), 4.9 (base)	
FSC threshold	0.143	0.143
Map resolution range (Å)	4.2  to > 10	4.6  to > 10
Refinement		
Initial model used	De novo modelling and	De novo modelling and
mittal model used	homology modelling	homology modelling
Model resolution (Å)	4.5 (top), 4.4 (middle),	4.6
Woder resolution (A)	4.9 (base)	4.0
FSC threshold	0.143	0.143
Model resolution range (Å)	n/a	n/a
Map sharpening B factor ( $Å^2$ )	-149 (consensus)	-213
Wap sharpening B factor (A)	-149 (conscilsus) -198 (top)	-213
	-190 (middle)	
	-177 (bottom)	
Model composition	-177 (bottom)	
Non-hydrogen atoms	15,309	12,424
Protein residues	3,827	3,106
Ligands	0	0
B factors ( $Å^2$ )	O	U
Protein	not estimated	not estimated
Ligand	not estimated	not estimated
R.m.s. deviations		
Bond lengths (Å)	0.23	0.22
Bond angles (°)	0.48	0.50
Validation	0.48	0.50
MolProbity score	1.44	1.12
Clashscore	3	3
Poor rotamers (%)	0	0
Ramachandran plot	V	V
Favored (%)	95.95	98.88
Allowed (%)	3.56	1.12
Disallowed (%)	0.49	0

# Extended Data Table 2 | Features of individual FA core complex subunits and structural models

Protein	Length (aa)	Sequence features	Models generated in this study	Maps used for modelling	Sequence identity / similarity between Gallus gallus and Homo sapiens (%)
FANCA	1,421	α-helical	N/A	N/A	49.1 / 65.6
FANCB	867	Possible $\beta$ -propeller, plus $\alpha$ -helical	De novo modelling of SSEs*	Focused map for top (EMD-10291), middle (EMD-10292), and base regions (EMD-10293), consensus map with T=5 (EMD-10290) and subcomplex map (EMD-10294)	44.1 / 63.2
FANCC	559	α-helical	SSEs placed in maps, not assigned	Focused map for base (EMD-10293)	49.0 / 65.1
FANCE	520	α-helical Crystal structure of human orthologue (C-terminal half; PDB 2ILR)	SSEs placed in maps, not assigned	Focused map for base (EMD-10293)	40.9 / 54.1
FANCF	350	α-helical Crystal structure of human orthologue (C-terminal half; PDB 2IQC)	C-terminal region from homology model based on PDB 2IQC	Focused map for base (EMD-10293)	38.5 / 51.6
FANCG	648	α-helical (TPR)	TPR domain from homology model (I- TASSER)	Focused map for top region (EMD-10291)	37.3 / 52.9
FANCL	373	ELF, URD and RING domains. Crystal structures of human (central domain, PDB 3ZQS; RING domain, PDB 4CCG) and <i>Drosophila</i> orthologues (full-length, PDB 3K1L)	ELF domain in FANCL <sub>top</sub> . ELF, URD and RING domains in FANCL <sub>base</sub> . Homology models based on PDB 3K1L	Focused map for top (EMD-10291) and base (EMD-10293)	69.9 / 82.7
FAAP100	888	Possible $\beta$ -propeller, plus $\alpha$ -helical	De novo modelling of SSEs	Focused map for top (EMD-10291), middle (EMD-10292), and base regions (EMD-10293), consensus map with T=5 (EMD-10290) and subcomplex map (EMD-10294)	55.6 / 68.7

SSE, secondary structure elements



Corresponding author(s):	Lori A Passmore
Last updated by author(s):	Sep 7, 2019

# **Reporting Summary**

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

$\overline{}$					
Ç	tっ	11	ist	т.	$\sim$
٠,	_		ורו		

For	all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.
n/a	Confirmed
	$oxed{oxed}$ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
	🔀 A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
$\boxtimes$	The statistical test(s) used AND whether they are one- or two-sided  Only common tests should be described solely by name; describe more complex techniques in the Methods section.
$\boxtimes$	A description of all covariates tested
$\times$	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
$\boxtimes$	For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i> ) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted <i>Give P values as exact values whenever suitable.</i>
$\boxtimes$	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
$\boxtimes$	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
$\boxtimes$	Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i> ), indicating how they were calculated
	Our web collection on <u>statistics for biologists</u> contains articles on many of the points above.
So	ftware and code

Policy information about availability of computer code

Data collection CryoEM data collected on Titan Krios microscope was performed with EPU (FEI/Thermo Fisher Scientific)

Data analysis Relion v2, Relion v3.0-beta, Eman2, MotionCor2, gCTF, ResMap, Coot, I-TASSER, Refmac, Marcoil, CCbuilder2.0, Xcalibur 2.2 (Thermo Fisher), Masslynx 4.2 (Waters), SUMMIT, Protein Lynx Global Server software (Waters), DynamX sotware (Waters), Xi software suite

(version 1.6.746) and XiFDR version 1.1.26.58, UCSF Chimera, ChimeraX, Phenix

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

### Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

CryoEM maps generated during this study have been deposited in the Electron Microscopy Data Bank (EMDB) with accession codes EMD-10290 (FA core complex consensus), EMD-10291 (focused classification top region), EMD-10292 (focused classification middle region), EMD-10293 (focused classification base region) and EMD-10294 (subcomplex). Models generated during this study have been deposited in the protein databank (PDB) with accession codes 6SRI (FA core complex) and 6SRS (subcomplex). Native MS data is available from figshare with accession code: 10.6084/m9.figshare.9692192. Crosslinking MS data has been deposited in the PRIDE database with accession code PXD014282. All other data are available from the authors upon reasonable request.

Field-spe	cific reporting		
Please select the or	e below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.		
Life sciences	Behavioural & social sciences Ecological, evolutionary & environmental sciences		
For a reference copy of the	e document with all sections, see <u>nature.com/documents/nr-reporting-summary-flat.pdf</u>		
Life scien	ces study design		
All studies must dis	lose on these points even when the disclosure is negative.		
Sample size	Sample sizes were chosen based on previous experience and published studies to evaluate reproducibility of assays. For cryo-EM, the initial number of particles was ~1,950,000, which was sufficient to obtain the stated resolution after 3D classification .		
Data exclusions	No data were excluded.		
Replication	All experiments (purifications, ubiquitination assays, pulldowns, nativeMS) were performed at least two or three times (exact number of replicates given in text). All attempts to replicate results were successful.		
Randomization	Randomization is not relevant to the experiments performed in this study.		
Blinding	Blinding is not relevant to the experiments performed in this study.		
We require informatic system or method list  Materials & exp n/a Involved in th  Antibodies  Eukaryotic  Palaeontolo Animals and	ChIP-seq  ell lines  gy  MRI-based neuroimaging  other organisms  arch participants		
Antibodies used	HA-probe (F-7) HRP monoclonal, Santa Cruz, Cat# sc-7392HRP, Lot# H3017, dilution 1:1000		
Validation	In this study, the HA antibody was used in Western blots to probe HA-tagged Ubiquitin. Western blot analysis of HA-tagged fusion proteins showing N-terminal HA-tagged JNK2 and JNK1 and C-terminal HA-tagged Daxx was performed by the manufacturer. In addition, we could verify the Western blotting results by Coomassie blue staining.		
Eukaryotic ce	ell lines		
Policy information about <u>cell lines</u>			
Cell line source(s)	Sf9, Oxford Expression Technologies Ltd, Cat No. 600100		

Authentication

Mycoplasma contamination

Commonly misidentified lines (See <u>ICLAC</u> register)

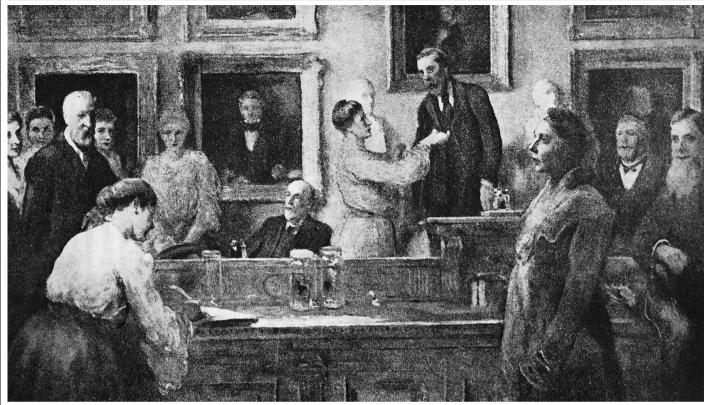
Cell line was not authenticated.

Cell line was negative for mycoplasma.

No commonly misidentified cell lines were used.

# Work





The Linnean Society of London first admitted women in 1905.

# CAREERS AND CONTROVERSY BEFORE THE FIRST WORLD WAR

For decades after Nature's launch in 1869, women's contributions to science were played down by both the journal and wider society. By Claire Jones

n its 150 years of existence, Nature has witnessed the emergence of science as a profession. But as research moved from a domestic to an institutional setting, women became increasingly invisible, and the historical narrative became resolutely male.

I aim to redress the balance by identifying the barriers that women faced and how they worked around them, gaining access to scientific education and chipping away at societies, journals and universities. Gradually, they widened the corridors of power for those who followed.

My focus is narrow - the United Kingdom in the late nineteenth and early twentieth centuries - but this was Nature's heartland in its first 50 years. And, for better or for worse, the British Empire provided a backdrop for scientific research in that era.

Wherever we look, women have been mostly absent from the story of science. To retrace the steps of these workaday women – not all heroines – of science is to understand how far we have travelled towards equity in the scientific workforce.

You could be forgiven for thinking that

"Acrimony was not unusual when the question of women's admission to societies was raised."

there was no such thing as a career in science for women before the mid-twentieth century. Our popular understanding of science as an essentially female-free zone for most of its existence is seldom challenged.

Yet women adopted various scientific guises before Nature was founded, and even occasionally appeared on its pages in its early years. This is not to say that science was a female-friendly career; serious prejudice and discrimination severely limited women's opportunities. However, recognizing the women who contributed to the enterprise despite these barriers debunks the myth that science was (and is) inherently male.

Early in the nineteenth century, women used spaces seen as more appropriately 'feminine' to

# **Work/Careers**

negotiate a way into science. Science writing, especially for children or popular audiences, scientific illustration and translation were all comfortable niches in which women could participate without threatening male pre-eminence or ideals of femininity.

Michael Faraday famously credited British science writer Jane Marcet's *Conversations on Chemistry* (1805) for inspiring him to take up science. Marianne North was a noted botanical illustrator, scientist and discoverer of plants. Later, astronomer Agnes Clerke negotiated a successful career as a writer of popular books on astronomy in the 1880s and 1890s, winning the Royal Institution's Actonian Prize in 1893.

### **Learned societies**

At the time of *Nature's* launch, most learned societies were male-only. In 1991, science historian Londa Schiebinger at Stanford University in California noted that for 300 years, the only permanent female presence at the Royal Society was a skeleton preserved in the anatomy cupboard<sup>1</sup>. In common with other elite scientific bodies, the society resisted admitting women as fellows until 1945, 26 years after the Sex Disqualification (Removal) Act 1919 was passed. Among other things, the act decreed that "a person shall not be disqualified by sex or marriage ... for admission to any incorporated society (whether incorporated by Royal Charter or otherwise)".

Nature was quick to rebuke the French Academy of Sciences² when it denied admission to physicist and chemist Marie Curie in 1911 – even though she had won a Nobel prize eight years previously. "It is incomprehensible ... on any ethical principles of rightness and justice," Nature wrote, "that because Curie happens to be a woman she should be denied the laurels which her pre-eminent scientific achievement has earned for her."

Women fought back, too. Around 1900, there was a concerted effort by a group led by evolutionary botanist Marian Farquharson, to gain admission to scientific societies. After strong debate between the fellows, 11 women were admitted to the Linnean Society in 1905. The society got its own back on Farquharson, however, by rejecting her application. She had to wait until 1908, when objections had died down, to be elected.

Acrimony was not unusual when the question of women's admission to societies was raised. When the Royal Geographical Society considered the issue in the decades around 1900, heated argument between fellows and members of the society's council broke out in the letters page of *The Times*. Exclusion from learned societies hindered women's access to networks, libraries, grants and collaboration, and made the career landscape very different for women than for men.

Why the raw antipathy to women? One reason was that science itself often taught ideas – now

discredited – that there were innate differences in intelligence between the sexes that would limit women's suitability for science. Darwin argued that evolutionary competition led to the higher development of male brains and of female emotions.

As a result, people saw the admission of women as threatening to dumb down proceedings and harm the status of elite

"Elite societies might have baulked at having female fellows, but women still managed to find a way in."

societies. Thomas Henry Huxley, a biologist and anthropologist who earned the sobriquet 'Darwin's bulldog' for his advocacy of evolution, worked to prevent women's admission to the Geological Society and the Ethnological Society of London, explicitly to preserve society status and prestige<sup>3</sup>. Ideologically informed theories of male and female brains

and resulting intellectual deficit are remarkably persistent, as neuroscientist Gina Rippon demonstrates in her 2019 book *The Gendered Brain*, which uses science to demolish these ideas. Rippon criticizes, in particular, modern evolutionary psychology and brain studies that look for differences between the sexes and, when they find it, consider only biological explanations.

However, the impact of these views – on women who were (and have been) internalizing them, and on the scientific community at large - cannot be ignored. Mathematician and astronomer Mary Somerville, widely celebrated in her time, remarked in entries in Personal Recollections, from Early Life to Old Age, of Mary Somerville, published posthumously in 1874, that she had "no originality ... that spark from heaven is not granted to the [female]". A review4 of her book in Nature identifies Somerville's genius as "wholly exceptional", because "women are not by nature adapted for studies which involve the higher processes of induction and analysis". Despite her unique scientific bent, the review takes pains to point out that



Elizabeth Brown was a founding member of the British Astronomical Association in 1890.

ROYAL ASTRON. SOC.



In the early 1900s, Marie Stopes received a grant from the Royal Society.

Somerville was still "beautifully womanly". Somerville had not only translated Pierre-Simon Laplace's notoriously difficult Traité de Mécanique Céleste (as Mechanism of the Heavens in 1831), she had also extended it with explanatory notes and her book was adopted as the standard text for higher mathematics at the University of Cambridge, UK. Indeed, the term'scientist' was coined for Somerville in the 1840s by Cambridge don William Whewell, as an alternative to 'natural philosopher' or 'man of science'.

GL ARCHIVE/ALAMY

Newer learned societies were not so choosy. These sprang up in large numbers towards the end of the nineteenth century as science specialized and associations emerged for amateur enthusiasts, teachers and women, Indeed, some women took key roles in these societies. For example, several were active in the British Astronomical Association, participating in expeditions, serving on its council and editing its journal. Elizabeth Brown was a founding member of the association: she headed the Solar Section of the Liverpool Astronomical Society, formed in 1881, which evolved into the British Astronomical Association in 1890.

Astronomy provided particular opportunity for women, arguably because practitioners remained in the field when other sciences professionalized and moved from the home to institutional spaces that excluded women. Botany, too, with its history as a feminized pursuit from the eighteenth century, proved welcoming, as did palaeobotany, which was strongly female-oriented in the first decades of the twentieth century5. Female palaeobotanists researching and publishing at this time include Margaret Benson at Royal Holloway College, University of London; Agnes Arber, who graduated from Newnham College in Cambridge; Henderina Scott, who researched and collaborated in a domestic setting; and Marie Stopes at the University of Manchester.

# **Collaboration sans compensation**

Elite societies might have baulked at having female fellows, but women still managed to find a way in, and participated in research in other ways, too. Between 1880 and 1914, some 60 women contributed to the Royal Society by authoring or co-authoring published papers or by demonstrating at the annual soirée, a highlight of the London social season that continues today<sup>6</sup>.

Some women, including palaeontologist Dorothea Bate and Stopes (who is best known for her later work on birth control and notorious for her later endorsement of eugenics), even received grants from the Royal Society to fund their research. Stopes' scientific career saw her travel widely for research, accept government commissions, publish nearly 40 scientific papers and produce important insights into coal-forest ecology. She earned doctorates from the University of Munich in Germany and from University College London, and became the first woman to join the science teaching staff at the University of Manchester.

Our modern understanding of a salaried science professional did not become completely valid until the second decade of the twentieth century, although men (and some women) did assume such roles from the 1870s onwards, often on the back of emerging technologies and industries, such as electrical engineering. Even when they had university training, women tended to secure low-status, routine roles such as research assistants and human

calculators at, for example, the Royal Observatory in Greenwich in the 1890s and at Imperial College London from its establishment in 1907.

However, it was far from unusual for women scientists to work alongside salaried men vet receive no remuneration for their labours. Bate, for instance, worked with the Natural History Museum in London from 1898, but was never paid and nor was she made a member of staff until 1948, when she was in her late 60s. The idea of a middle-class woman receiving payment violated all ideals of respectable femininity.

Earlier in the century, this concept also affected Eleanor Ormerod, who provided economic advice on agricultural problems and pests. It was easier for a middle-class woman of means to carry out research or to do so alongside teaching, one of the few respectable careers for women. However, working-class women could find a pathway into science from a business direction. Nautical-instrument maker, inventor and navigation writer Janet Taylor ran a nautical academy in the East End of London in the 1860s and 1870s, with the Admiralty as one of her clients.

Ormerod was a pioneering technological scientist who was instrumental in establishing the discipline of economic entomology in Britain, in particular through her annual reports published from 1877 to 1901. Although Ormerod was self-taught and possessed no formal qualifications - something not unusual for women or men at the time, given the amateur tradition in science - she advised and lectured on training at various colleges and was an examiner at the University of Edinburgh, UK.

Ormerod also participated in international collaborative research, acted as an expert witness in legal cases and was commissioned as a consultant entomologist to the Royal Agricultural Society in 1882. However, she was not paid, and received only occasional expenses, despite giving her expertise for free for the next ten vears.

One route into science for women at this time was through collaboration with a husband or other male family member. Yet, even for the most egalitarian of scientific partnerships, it was the man who tended to get the kudos, with his female collaborator cast in the role of

Many women accepted this. Two examples are astronomer Margaret Huggins and Scott, a pioneer slow-motion filmmaker, botanist and palaeobotanist. Both women were independent researchers, but bought into the era's perceptions about wives being 'helpmeets' to their husbands.

Yet Scott's husband was a strong supporter of women scientists, unlike Huggins's, who complained that illness had prevented him blocking the award of the Royal Society Hughes Medal for original research to electrical engineer and physicist Hertha Ayrton in 1906. When

Ayrton died in 1923, an obituary in Nature asserted that, instead of pursuing her own scientific interests, she should have looked after her husband, and "put him in carpet slippers when he came home", so that he could have better devoted his efforts to his scientific work7. Avrton might have succeeded as a scientist but. according to her obituarist at least, she did not succeed as a wife.

Some of the research for which Ayrton was honoured had been done in her husband's laboratories at the Central Institution in Kensington, London. This included work on her book The Electric Arc (1902) which became the go-to resource on the subject and had been serialized in Nature in 1899.

When her husband died, Ayrton lost access to this institutional space and so turned her living room into a laboratory. Her confinement to the domestic sphere at a time when emphasis was being placed on precise measurements and instrumentation prompted questions over her research and the credibility of her science.

Women had to tread particularly carefully when they entered the laboratory, which was seen as a space for masculine display. Women's presence there could prompt scepticism, if not outright hostility, especially when access was for research rather than educational purposes. This antagonism often led to the development of parallel facilities, such as the Balfour Biological Laboratory for Women at the University of Cambridge in 1884.

As the new century approached, more women were accessing a university education in science, and the idea of a professional female researcher was no longer an oddity. The University of London was a key player here, opening up its degrees to women and men on an equal basis (except for medicine) from 1878.

Science was particularly strong at London's Royal Holloway and Bedford women's colleges. When Royal Holloway opened its doors in 1886, it did so with well-equipped chemical and biological laboratories.

Women were allowed to graduate from Scottish universities after the passing of a special act in 1889 (apart from degrees in medicine, which were not conferred on women until 1916).

But the battle for women's higher education was not wholly won. That year, physician William Withers Moore used an address to the British Medical Association to warn against university education for women owing to the "dangers" it posed to female reproductive health and mental well-being.

# "A more acceptable route into science was teaching in one of the colleges or high schools for girls."

Undaunted by his warnings, some women graduates began to take on research posts and embark on higher degrees in the United Kingdom, Germany and the United States. For example, mathematician and biostatistician Karl Pearson employed a number of women at Galton Laboratory, established in 1904 at University College London. Alice Lee, who had studied mathematics at Bedford College, went on to become a doctor of science under

his supervision. Women were not awarded degrees at Cambridge until 1948 (27 years after Oxford began conferring them), but they did study natural sciences and made contributions to research. Between 1902 and 1910, female researchers at Newnham College were instrumental in founding the science of genetics8, working alongside biologist William Bateson.

A more acceptable route into science was teaching in one of the colleges or high schools for girls that were being established at the end of the century. Many of the female graduates found their scientific niche in teaching, including Cambridge mathematician Sara Burstall. who became head of Manchester High School for Girls in 1898.

However, not everyone was pleased with this development. Chemist William Armstrong used his report for the 1904 Mosely Education Commission to emphasize the "mental disabilities" that evolution had bestowed on women and to issue dire warnings about the "ruinous" effects of allowing them to "contaminate" boys by teaching them science.

The important work of female scientists during the First World War – stepping up to run laboratories while men were away at the front - is only just now being given due credit<sup>9</sup>. Stopes was recruited to the war effort by the UK government's Industrial Research Department, where she collaborated on research into the constituents of coal, Hilda Phoebe Hudson, like other female mathematicians, joined the Air Ministry to research problems in aeronautical engineering.

The popular history of women in science tends to celebrate romantic 'heroines' such as Ada Lovelace (who, later in her short life at least, used her mathematical prowess mostly to gamble) or two-time Nobel-prizewinning Curie, rather than the workaday women who made their way in science as best they could often very successfully.

Remembering the breadth of female participation will not only end science's 'disappearing woman' trick, it might also illuminate the current gender imbalance by making the point that science is, and always has been, for women as much as for men.

Claire Jones is a historian of science and senior lecturer at the University of Liverpool,

- Schiebinger, L. in The Mind has No Sex? Women in the Origins of Modern Science 26 (Harvard Univ. Press. 1991).
- Nature 85, 342 (1911).
- Richards, E. in Victorian Science in Context (ed. Lightman, B.) 126 (Univ. Chicago Press, 1997)."
- Nature 9. 417-418 (1874).
- Fraser, H. E. & Cleal, C. J. in The Role of Women in the History of Geology (eds Burek, C. & Higgs, B.) 51-82 (Geological Society, 2007).
- Jones, C. in Femininity, Mathematics and Science, 1880-1914 177-184 (Palgrave Macmillan, 2009).
- Armstrong, H. E. Nature 112, 800-801 (1923)
- Richmond, M. L. Isis 92, 55-90 (2001)
- Fara, P. A Lab of One's Own: Science and Suffrage in the First World War (Oxford Univ. Press, 2018)



Botanical illustrations by Marianne North made their mark in the mid-nineteenth century.



Crystallographer Kathleen Lonsdale was one of the first two women to be elected as a fellow of the UK Royal Society, in 1945.

# THE WOMEN WHO CRACKED THE GLASS CEILING

After the First World War, female scientists gained footholds in academia as well as industrial and government research, despite facing prejudice and many other barriers. By Sally Horrocks

cientific career opportunities saw a boost during the First World War as a result of the realignment of science to the military. For the first time, scientists worked on problems ranging from aviation and submarine detection to chemical warfare. After the war, this expansion continued, particularly in industry. Biochemist Kathleen Culhane Lathbury was one female scientist who benefited from that. During the 1920s and early 1930s, she worked for British Drug Houses, one of the leading pharmaceutical firms in the United Kingdom, which I focus on here. In her post, Lathbury oversaw insulin manufacturing.

But because the drug maker's dining room

was male-only, she was excluded from the social interactions that happen when dining with colleagues. In notes for a talk that she gave on women in the chemical industry, Lathbury said that the male graduate "is usually given quite a dignified position from the beginning. The girl who worked side by side with him at the university is hard up and constantly humiliated ... Even if her work is intellectually satisfying, she will be expected to attain results from the ground floor for which her male equivalent is given the help of a little altitude."

In my role as a science historian, since 2011, I have been senior academic adviser to An Oral History of British Science, a National Life Stories project in collaboration with the British Library. The project has collected memories of the lives and careers of British scientists since the 1940s. (Edited extracts are available at www.bl.uk/ voices-of-science, with full interviews accessible at sounds.bl.uk/oral-history/science.)

In 1922, Lathbury graduated from Royal Holloway College in London with a chemistry degree. She signed her job applications 'K. Culhane' to mask her gender, and worked for no pay at the Royal Institute of Chemistry, concluding that "for women in the chemical industry, magnificent health and a thick skin are more important than a knowledge of chemistry".

As her story demonstrates, the inter-war period was one of increased employment

# **Work/Careers**

of women in science, but also of continued exclusion and segregation. After the First World War, what had been wartime research organizations grew, while those established before 1914, including corporate laboratories that had existed since the early 1890s, consolidated their positions, contributing to the growth of a new, technical middle class. But the career patterns of female scientists differed greatly from those of their male counterparts, and the disparity has persisted, even during the Second World War and the first few decades of the cold war.

In the United Kingdom — which I focus on here — women were also limited by an expectation that they would resign from work once they married. In some cases, including in the civil service, such resignation was a formal requirement with limited exceptions, so many of the women who enjoyed lengthy careers at government research organizations remained single. Women in the civil service could be exempted from this bar if their work was deemed to be of sufficient national importance, but, in practice, very few actually received exemptions.

One example of the paucity of exceptions was aeronautical engineering researcher Frances Bradfield, who studied mathematics and physics at Newnham College, Cambridge (a women's college established in 1871). She joined the UK government's Royal Aircraft Establishment (RAE) in Farnborough in 1918, along with fellow Newnham graduate Muriel Barker.

Bradfield remained at the RAE until her retirement in 1955, taking charge of small wind tunnels, mentoring many of her younger male colleagues and gaining the respect of her peers. Barker married colleague Hermann Glauert in 1922, and left her post.

Fellow Farnborough employee Beatrice Shilling, an expert on aero-engines, however, was one of the few who received an exemption when she married RAE mathematician George Naylor in 1938, leaving the RAE only when she retired in 1969. Shilling developed a device to counter engine cut-out in early Spitfire and Hurricane planes during the Battle of Britain in 1940.

### Marriage and mobility

In 1945, X-ray crystallographer Kathleen Lonsdale (née Yardley) and biochemist Marjory Stephenson became the first two women to be elected fellows of the Royal Society, the United Kingdom's national academy of sciences. Stephenson, who was employed for much of her career by the Medical Research Council, had won her first university appointment in 1943.

Physics Nobel laureate William Henry Bragg had supported Lonsdale in her career at University College London and at the Royal Institution in London. Lonsdale worked from home after starting a family in 1929, and her husband assumed domestic responsibilities. A pacifist and penal reformer, Lonsdale served a month's sentence in London's Holloway Prison during



Stephanie Shirley built computers at the Post Office Research Station in the 1950s.

the Second World War because, as a Quaker, she refused to register for civil-defence duties.

Beryl Platt, by contrast, studied engineering at the University of Cambridge, UK, and joined the Hawker Aircraft Company in 1943. Platt had switched from mathematics to mechanical engineering (as one of 5 female students alongside 250 male undergraduates) when she arrived at Girton College in Cambridge two years earlier, because the UK government offered a state bursary to encourage engineering undergraduates as part of the war effort. After a brief post-war career in air safety for British European Airways, she ended her professional career in engineering when she married textiles manufacturer Stewart Platt in 1949.

Women who married fellow scientists, particularly those who worked in universities, were sometimes able to continue their involvement in research. Organic chemist Gertrude Robinson, who earned a master's degree in 1908, worked at the University of Manchester as a research assistant to Chaim Weizmann (who became Israel's first president in 1949), before marrying future Nobel laureate Robert Robinson in 1912. She collaborated with him on research in organic chemistry, publishing more than 30 papers. The couple spent a brief period at the University of Sydney in Australia,

one of the growing number of universities in the English-speaking world that recruited UK academic researchers and staff.

Such international mobility was a feature of professional scientific careers from the nineteenth century onwards, but men were more likely to take advantage of it than were women. More than 16% of UK-born chemists who joined the Royal Institute of Chemistry between 1887 and 1943 worked overseas at some point during their careers.

### War work

As the world pivoted towards the Second World War in 1939, the United Kingdom started to see scientists as a national asset, and the Ministry of Labour and National Service established procedures for recruiting and training scientists and engineers. Men who were qualified to embark on courses in the physical sciences or in engineering were exempt from the armed services while they completed their degrees. These were compressed from three to two years, even in Scotland, where honours degrees typically last for four years. But the ministry actively discouraged universities from increasing the proportion of female students in science and engineering, despite the nation's demand for expertise.

Both women and men were directed into war work after completing their studies, however. Some were roped in even earlier. For example, microbiologist Nada Jennett (née Phillips) and fellow University of Bristol students spent one of their holidays working for pharmaceutical company Glaxo on penicillin production problems.

After the war, Jennett trained as a teacher and worked in laboratories at the university and in a hospital in Cardiff until her first child was born. She taught science part-time before returning to microbiology, then developed a second career lecturing in garden design.

For men, wartime work was often the foundation of long and successful careers, but for women it generally represented a short interlude before full-time domestic responsibilities, which might be followed by unpaid voluntary work or by part-time paid employment, but rarely a permanent post. Some employers who had been reluctant to hire women relented, among them Imperial Chemical Industries (ICI), then Britain's largest chemical manufacturer.

ICI advertisements specified a preference "for women chemists of British nationality", perhaps helping to explain why refugee women who were scientists were not always able to find relevant work, even if they had impressive qualifications. In March 1941, for example, the journal Chemistry and Industry carried this advert: "LADY CHEMIST. German Refugee, aged 37. PhD (Berlin), seeks a position. Some research experience in Rubber Chemistry and accustomed to conduct searches in libraries and translate from German and French."

Women who were married, had children and had left science to concentrate on domestic responsibilities but wanted to contribute to the war effort also found suitable work hard to come by, Lathbury, for one, ended up working in statistical quality control at the Royal Ordnance Factory after a brief stint as a wages clerk.

In 1939, Joan Strothers and Sam Curran, then physics PhD students at the Cavendish Labo-

# "For many women, continuing to work after marriage was often the only practical option."

ratory in Cambridge, were trying to develop a proximity fuse, an explosives detonator that triggered only when near the target. They married a year later and moved to the Telecommunications Research Establishment, where Curran worked on centrimetric radar systems  $for installation in aircraft, while Strothers \, was \,$ part of the countermeasures group. Here, she developed the idea that led to Operation Window – the scattering of strips of metallic foil from aircraft to deceive enemy radar, a technique that was successfully used on D-Day.

# **Expanding opportunities**

Towards the end of the Second World War. workforce planners expected a contraction in military research, enabling UK industry to recruit researchers to help recover the economy after the war. But this contraction proved short-lived. Defence research, including work on a British atomic-bomb project, rapidly expanded in the late 1940s and early 1950s, creating many new jobs in research organizations.

Asmall but growing number of graduate-level female scientists found employment in defence-research establishments and, thanks to the 1946 abolition of the marriage bar in the civil service, could now continue their careers after marriage.

However, without maternity-leave legislation or a provision for childcare, many married women could not continue to work. And, although some enjoyed long careers, few reached senior positions.

An exception was the naval engineer Elizabeth Killick, whose career began in the early 1950s, Killick, who died in July 2019 aged 94, became deputy chief scientific officer and head of the Weapons Department at the Admiralty Underwater Weapons Establishment. In 1982, she also became the first woman to be elected to what is now the Royal Academy of Engineering.

Expanded UK government support for health, education, employment and social security after the Second World War also generated new opportunities for scientists, including posts in biological sciences, which tended to be popular with female researchers. Organizations such as the UK Public Health Laboratory Service and the advice services coordinated by what was then the Ministry of Agriculture, Fisheries and Food also employed women.

Measures agreed in 1955 meant that from 1960, women who worked for the state received the same wages as men. For women, this made careers in government research and universities more attractive than those in industry, in which differential pay rates and benefits remained the norm.

But even after the lifting of the marriage bar, women who did secure permanent academic posts often had to assume significant teaching and administrative burdens while their male colleagues were free to focus on research -work that brought greater prestige and faster promotion.

In 1947, for example, Florence R. Shaw was appointed to an assistant lectureship at University College, Leicester (now the University of Leicester), and was promoted to lecturer in 1948. But she published little after being elected a fellow of the Royal Institute of Chemistry in 1949 and, on her retirement in 1965, was praised for her teaching contribution as "a loyal and steadfast colleague in the Chemistry Department, to whom many of our graduates owe a great deal".

Female researchers who pursued scientific careers during the post-war period faced emotional and practical challenges in the predominantly male environments. Many experienced self-doubt and had to come up with strategies to improve their status without seeming to be openly confrontational.

Stephanie Shirley, who arrived in the United Kingdom in 1939 as a refugee from Nazi Germany, worked at the Post Office Research Station in the 1950s, building computers from scratch. She recalls, "If you're the only one, if you fail, you fail for all women, and they say,



Engineer Beryl Platt (left) with an associate on the occasion of his wedding.

# Work / Careers



In the 1940s, Beatrice Shilling developed a device to stop aeroplane engines cutting out.

'Well, we tried one of those and she was awful.' Whereas if you succeed, it's also remembered, but somehow the presumption is that, 'We had her and she was good; at least we'll try another one and see if it works again."

In the late 1960s, recognition of the barriers to women's access to scientific careers began to grow. These obstacles came to be seen as problems that needed to be addressed, rather than as the inevitable consequences of women's prioritizing of family obligations over career aspirations. From the 1970s, many of these formal barriers were removed. Female scientists in the United Kingdom and elsewhere benefited from legislative changes that promoted greater equality in employment and provided for maternity leave.

The three key pieces of UK legislation were the Equal Pay Act (1970); the Sex Discrimination Act (1975), which outlawed discrimination in employment on the grounds of gender or marital status; and the Employment Protection Act (1975), which established the principle of paid maternity leave, although it did not initially cover all women.

In the United States, Title IX of the Education Amendments Act (1972) outlawed discrimination on the basis of gender in education or activities receiving federal funding. But as Margaret Rossiter showed in the 2012 third volume of her book *Women Scientists in America*,

those researchers had to fight hard to ensure it was implemented.

At a global level, the United Nations decreed 1975 to be International Women's Year, and the first UN Conference on Women was held that year in Mexico City. In 1979, the UN

# "Greater diversity in the workforce came to be seen as an economic asset."

Convention on the Elimination of All Forms of Discrimination Against Women was adopted.

The European Economic Community (from 1993, the European Community; from 2009, the European Union) was also a powerful force for promoting equality legislation in its member states, including the extension of maternity leave to all working women in the United Kingdom in 1993 and the extension of paternity leave in 2010. (Paid paternity leave was introduced in the United Kingdom in 2003.)

Legislative change and international conventions did not mean, however, that the expectations of employers or female scientists themselves changed suddenly, or that discrimination disappeared overnight.

Meteorologist Julia Slingo, whose first daughter was born in 1980, opted to leave

her job at the UK Met Office rather than take maternity leave. She returned to work in 1981 after being offered flexible working arrangements, an option she continued to take advantage of even after she accepted a new role in the United States in 1986. She later returned to full-time work and enjoyed a successful career before retiring in 2016 as chief scientist of the Met Office, a year after she was elected a fellow of the Royal Society.

Such flexible arrangements became more widely available from the 1990s. This was because greater diversity in the workforce came to be seen as an economic asset, making gender equality a matter of sound business practice rather than merely about the pursuit of social justice.

This business-case approach has also prompted efforts in Europe and the United States to address other aspects of diversity, including factors such as ethnicity, disability, sexual orientation and socio-economic status. Such an approach tends to focus on providing equality of opportunity to existing educational and employment structures rather than — as feminist critics have been advocating since at least the 1990s — on challenging the imbalances of power that form the basis of under-representation.

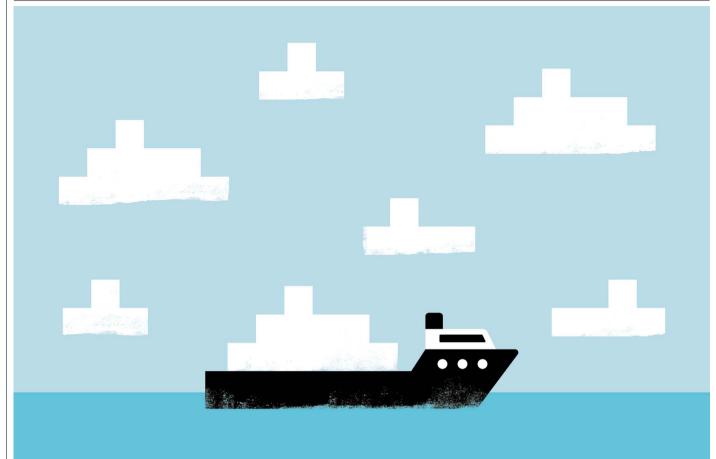
British female scientists who started their careers in the years after the First World War were a small minority in a relatively new profession that was concentrated in Europe and North America and was only just beginning to emerge elsewhere.

Their counterparts in the twenty-first century are members of a global community of nearly 8 million researchers. More than 40% of those are in Asia, although the proportion of female researchers worldwide is less than 30%. Whereas many of the formal barriers to women's participation in UK science that existed in 1919 disappeared in the twentieth century, many fields continue to be numerically and structurally male. In these areas, career progress for women — as was the case a century ago — involves a challenging process of trying to work in male-oriented environments while seeking to maintain their own gender identities.

Female scientists might no longer be forced to choose between career or marriage and family. But they continue to face many challenges, with workplace cultures and reward structures still designed mainly to accommodate male-oriented norms and career paths.

Sally Horrocks is associate professor of contemporary British history at the University of Leicester, UK. She thanks members of the An Oral History of British Science team past and present, as well as Liz Bruton (Science Museum, London) and Graeme Gooday (University of Leeds, UK) for advice and encouragement.

e-mail: smh4@leicester.ac.uk



# **CONTAINERS** IN THE CLOUD

# Standardized platforms allow researchers to run each other's software – no installation required. By Jeffrey M. Perkel

urphy's law for the digital age: anything that can go wrong, will go wrong during a live demonstration. For Ben Marwick, that happened in front of a roomful of landscape-archaeology students in Berlin. The topic: computational reproducibility using Docker.

Docker is a software tool that generates 'containers' - standardized computational environments that can be shared and reused. Containers ensure that computational analyses always run on the same underlying infrastructure, fostering reproducibility. Docker thereby insulates researchers from the challenges of installing and updating research software. However, it can be difficult to use.

Marwick, an archaeologist at the University of Washington in Seattle, had become proficient in migrating Docker configuration files ('Dockerfiles') from one project to the next, making minor tweaks and getting them to work. Colleagues in Germany invited him to teach their students how to follow suit. But because every student had a slightly different set of hardware and software installed, each one required a customized configuration. The demo "was a complete disaster", Marwick says.

Today, a growing collection of services allows researchers to sidestep such confusion. Using these services – which include Binder, Code Ocean, Colaboratory, Gigantum and Nextjournal – researchers can run code in the cloud without needing to install more software. They can lock down their software configurations, migrate those environments from laptops to high-performance computing

clusters and share them with colleagues. Educators can create and share course materials with students, and journals can improve the reproducibility of results in published articles. It's never been easier to understand, evaluate, adopt and adapt the computational methods on which modern science depends.

William Coon, a sleep researcher at Harvard Medical School in Boston, Massachusetts, spent weeks writing and debugging an algorithm, only to discover that a colleague's containerized code could have saved a lot of time. "I could have just gotten up and running, using all of the debugging work that he had already done, at the click of a button," he says.

Scientific software often requires installing, navigating and troubleshooting a byzantine network of computational 'dependencies'

# Work/Technology&tools

- the code libraries and tools on which each software module relies. Some have to be compiled from source code or configured just so, and an installation that should take a few minutes can degenerate into a frustrating online odyssey through websites such as Stack Overflow and GitHub. "One of the hardest parts of reproducibility is getting your computer set up in exactly the same way as somebody else's computer is set up. That is just ridiculously difficult," says Kirstie Whitaker, a neuroscientist at the Alan Turing Institute in London.

### **Easier evaluation**

Docker reduces that to a single command. "Docker really provides reduced friction for that stage of the cycle of reproducing somebody else's work, in which you have to build the software from source and combine it with other external libraries," says Lorena Barba, a mechanical and aerospace engineer at George Washington University in Washington DC. "It facilitates that part, making it less error-prone, making it less onerous in researcher time."

Barba's team does most of its work in Docker containers. But that is a computationally savvy research group; others might find the process daunting. A text-based 'command-line' application, Docker has dozens of options, and building a working Dockerfile can be an exercise in frustration.

That's where the cloud-based services come in. Binder is an open-source project that allows users to test-drive computational notebooks — documents such as Jupyter or R Markdown notebooks, which blend code, figures and text. Colaboratory (free), Code Ocean, Gigantum and Nextjournal (the latter three have free and paid tiers) let users write code in the cloud as well and, in some cases, bundle it with the data to be processed. These platforms also allow users to modify the code and apply it to other data sets, and provide version-control features for reviewing changes.

Such tools make it easier for researchers to evaluate their colleagues' work. "With Binder, you have taken that barrier [of software installation] away," says Karthik Ram, a computational ecologist at the University of California, Berkeley. "If I can click that button, be dropped into a notebook where everything is installed, the environment is exactly the way you intended it to be, then you've made my life easier to go take a look and give you feedback."

Identifying required dependencies, and where to find them, varies with the platform. On Code Ocean and Gigantum, it's a point-and-click operation, whereas Binder requires a list of dependencies in a Github respository. Whitaker's advice: codify your computing environment as early as possible in a project, and stick with it. "If you try and do it at the end, then you are basically doing archaeology on your code, and it's really, really hard," she says. Ram developed a tool called Holepunch for

projects that use the statistical programming language R. Holepunch distils the process of setting up Binder into four simple commands. (See examples of our code running on all five platforms at go.nature.com/2ps9se1.)

The easiest way to try Binder is at mybinder.org, a free, albeit computationally limited, website. Or, for greater power and security, researchers can build private 'BinderHubs' instead. The Alan Turing Institute has two, including one called Hub23 (a reference to Hut 23 at the Second World War code-breaking facility at Bletchley Park, UK), that provides

"Researchers can be confident that their code will remain usable, whichever platform they choose."

greater computational resources and the ability to work with data sets that cannot be publicly shared, Whitaker says. The Pangeo community, which promotes open, reproducible and scalable geoscience, built a dedicated BinderHub so that researchers can explore climate-modelling and satellite data sets that can amount to tens of terabytes, says Joe Hamman, a computational hydroclimatologist at the National Center for Atmospheric Research in Boulder, Colorado. (Whitaker's team has published a tutorial on building a BinderHub at go.nature.com/349jscv.)

### Languages and clouds

Google's Colaboratory is basically a cross between a Jupyter notebook and Google Docs, meaning users can share, comment on and jointly edit notebooks, which are stored on Google Drive. Users execute their code in the Google cloud – only the Python language is officially supported – on a standard central processing unit (CPU), a graphics processing unit (GPU) or a tensor processing unit (TPU), a specialized chip optimized for Google's TensorFlow deep-learning software. "You can open up your notebook or someone else's notebook from GitHub, start playing around with it and then save your copy on Google Drive and work on it later," says Jake VanderPlas, a member of the Colaboratory team at Google in Seattle.

Nextjournal supports notebooks written in Python, R, Julia, Bash and Clojure, with more languages in development. According to Martin Kavalar, chief executive of Nextjournal, which is based in Berlin, the company has registered nearly 3,000 users since it launched the platform on 8 May.

Gigantum, a beta version of which launched last year, features a browser-based client that users can install on their own system or remotely, for cloud-based coding and execution in the Jupyter and RStudio coding environments. Coon, who uses Gigantum to run

machine-learning algorithms in the Amazon cloud, says the service makes it easy for collaborators to hit the ground running. "[They] can read through my Gigantum notebooks and use this cloud-compute infrastructure to do the training and learning," he explains.

Then there's Code Ocean, which supports both notebooks and conventional scripts in Python, R, Julia, Matlab and C, among other languages. Several journals now use Code Ocean for peer review and to promote computational reproducibility, including titles from Taylor & Francis, De Gruyter and SPIE. In 2018. Nature Biotechnology. Nature Machine Intelligence and Nature Methods launched a pilot programme to use Code Ocean for peer review; Nature, Nature Protocols and BMC Bioinformatics subsequently joined the trial. More than 95 papers have now been involved in the trial, according to Erika Pastrana, editorial director of Nature Research's applied-science and chemistry journals, and more than 20 of those have been published.

Felicity Allen, a computer scientist at the Wellcome Sanger Institute in Hinxton, UK, co-authored one study in that trial, which analysed the types of mutation that can arise from CRISPR-based gene editing (F. Allen *et al. Nature Biotechnol.* 37, 64–72; 2019). She estimates that it took a week to get the Code Ocean environment working. "The reviewers seemed to really like it," Allen says. "And I think it was really nice that it made an example that someone could just press 'go' on and it would run."

Although some worry about the long-term viability of commercial container-computing services, researchers do have options. Simon Adar, chief executive of Code Ocean, notes that Code Ocean 'compute capsules' are archived by the CLOCKSS project, which preserves digital copies of online scientific literature. And Code Ocean, Gigantum and Nextjournal allow Dockerfiles to be exported for use on other platforms. All of which means that researchers can be confident that their code will remain usable, whichever platform they choose.

Benjamin Haibe-Kains, a computational pharmacogenomics researcher at the Princess Margaret Cancer Centre in Toronto, Canada, adopted Code Ocean to respond quickly to critiques of an analysis he published in Nature (B. Haibe-Kains et al. Nature 504, 389-393; 2013). For him, Code Ocean provides a way to ensure his code can be used and evaluated by his team, peer reviewers and the broader scientific community. "It's not so much that an analysis must be correct or wrong," he says. "Nothing is really fully correct in this world. However, if you're very transparent about it, you can always communicate efficiently in the face of criticism. You have nothing to hide; everything is there."

**Jeffrey M. Perkel** is technology editor at *Nature*.

# The back page



# Where I work Terri Adams

've been a scientific glassblower for 33 years. For much of that, I've worked at the University of Oxford, where I design and create glass equipment that scientists can use for their research.

The piece I'm most proud of is a perfusion apparatus that is used to keep human organs functioning outside the body. But I also make glassware that is used throughout the university, such as high-vacuum manifolds, which are a series of knobs that operate a vacuum; glass apparatus for distillation and sublimation experiments; vessels with water jackets used to heat and cool materials; and high-temperature furnace tubes made of quartz or ceramic.

My workbench hosts an array of tools for working with glass, many of which were custom-made for specific jobs. Each tool reminds me of what I first used it for and makes me consider how I might use it again.

Most are made of carbon, and need to be highly polished before use because any irregularities will be transferred to the finish on the glass.

My workbench is also where I ponder the

design of new glassware. It's quite easy to sketch something on a piece of paper, but reproducing that concept as a workable piece of glass equipment is a much more difficult endeavour.

I find that a lot of my work relies on intuition: I instinctively know when the glass is the right temperature, or at what speed it needs to rotate on the lathe. Usually, I can tell when it's turning fast enough by the sound of the lathe.

Glassblowing is a declining art – worldwide, there aren't many schools that teach it any more. A lot of the work can be done by a computer, and there are now alternative materials to non-magnetic glassware.

However, I learn something new almost every week, and am inspired by knowing that a little piece of glassware that I've made has contributed in some way to the bigger picture of science when a researcher achieves milestone results.

**Terri Adams** is a professional glassblower at the University of Oxford, UK. **Interview by Sarah Boon.** 

Photographed for *Nature* by Leonora Saunders